SYSTEMS ENGINEERING
Research Center

# Defining the Meaning of a Major Modeling and Simulation Change as Applied to Accreditation

## A013 - Final Technical Report SERC-2012-TR-029

December 12, 2012


Principal Investigator

Mikel D. Petty, Ph.D., University of Alabama in Huntsville

Team Members

Philip W. Alldredge, University of Alabama in Huntsville

J. Cameron Beach, University of Alabama in Huntsville

Wesley N. Colley, Ph.D., University of Alabama in Huntsville

**Center for Modeling, Simulation, and Analysis**
University of Alabama in Huntsville
301 Sparkman Drive, Shelby Center 144
Huntsville AL 35899




Contract Number: H98230-08-D-0171

TO 0019, RT 038

**Report No. SERC-2012-TR-029**
**December 12, 2012**
UNCLASSIFIED

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **12 DEC 2012** | | **00-00-2012 to 00-00-2012** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Defining the Meaning of a Major Modeling and Simulation Change as Applied to Accreditation** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Center for Modeling, Simulation, and Analysis,University of Alabama in Huntsville,301 Sparkman Drive, Shelby Center 144,Huntsville,AL,35899** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

**Increasingly, decisions are being made based partially or entirely on models. These decisions may be quite important, with the potential of serious or unacceptable consequences if an incorrect decision is made. The models used to support the decisions may have been newly developed or modified versions of existing models. In the former case, it seems clear that a new model should undergo validation before it is used for any significant application. In the latter case, the matter may be less clear; must the modified model, which presumably was validated when it was initially developed, be subject to another round of validation due to the modifications? Several existing methods address the re-validation question, including methods developed by the Johns Hopkins University Applied Physics Laboratory, the Institute of Electrical and Electronics Engineers, the Joint Accreditation Support Agency, and the project?s sponsoring agency. These existing methods vary widely in several respects including level of detail, specificity with respect to modeling and simulation, degree of quantitativeness, ease of use, and applicability to the sponsoring agency. All include some form of the notion of risk, which is conventionally defined as the product of the likelihood of an incorrect decision and the consequences of such a decision. A new method, the Quantitative-to-Qualitative Risk-based (QQR) method, was developed to make a quantitative recommendation regarding the re-validation of a modified model. The QQR method was developed with these goals in mind: to be quantitative, repeatable, and transparent to consider both model modifications and model use risk; to focus on model types and simulation applications of interest to the sponsoring agency; and to be simple and accessible so as to encourage its use in practical applications. The QQR method estimates the probability that not re-validating a modified model will lead to unacceptable consequences, given the modifications made to it. That estimated probability is meant to be interpreted, in the context of a re-validation decision, as a quantitative recommendation to re-validate the model. The method is based on a conditional probability formula that separates the various parts of the estimated probability into distinct terms and factors, and it provides procedures for estimating each term and factor. A central feature of the QQR method is a missions-means decomposition that allows both the method?s user to precisely and effectively identify the nature and extent of the modifications that were made to the model and the method to consider those modifications in its estimate. The QQR method was validated using a set of re-validation scenarios, each describing a model the modifications made to it, and the decision to be based on the model. The QQR method?s revalidation recommendations for the scenarios were compared to those of a set of human experts**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| **unclassified** | **unclassified** | **unclassified** | **Same as Report (SAR)** | **62** | |

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense through the Systems Engineering Research Center (SERC) under Contract H98230-08-D-0171. SERC is a federally funded University Affiliated Research Center managed by Stevens Institute of Technology

## Abstract

Increasingly, decisions are being made based partially or entirely on models. These decisions may be quite important, with the potential of serious or unacceptable consequences if an incorrect decision is made. The models used to support the decisions may have been newly developed or modified versions of existing models. In the former case, it seems clear that a new model should undergo validation before it is used for any significant application. In the latter case, the matter may be less clear; must the modified model, which presumably was validated when it was initially developed, be subject to another round of validation due to the modifications?

Several existing methods address the re-validation question, including methods developed by the Johns Hopkins University Applied Physics Laboratory, the Institute of Electrical and Electronics Engineers, the Joint Accreditation Support Agency, and the project's sponsoring agency. These existing methods vary widely in several respects including level of detail, specificity with respect to modeling and simulation, degree of quantitativeness, ease of use, and applicability to the sponsoring agency. All include some form of the notion of risk, which is conventionally defined as the product of the likelihood of an incorrect decision and the consequences of such a decision.

A new method, the Quantitative-to-Qualitative Risk-based (QQR) method, was developed to make a quantitative recommendation regarding the re-validation of a modified model. The QQR method was developed with these goals in mind: to be quantitative, repeatable, and transparent; to consider both model modifications and model use risk; to focus on model types and simulation applications of interest to the sponsoring agency; and to be simple and accessible so as to encourage its use in practical applications.

The QQR method estimates the probability that not re-validating a modified model will lead to unacceptable consequences, given the modifications made to it. That estimated probability is meant to be interpreted, in the context of a re-validation decision, as a quantitative recommendation to re-validate the model. The method is based on a conditional probability formula that separates the various parts of the estimated probability into distinct terms and factors, and it provides procedures for estimating each term and factor. A central feature of the QQR method is a missions-means decomposition that allows both the method's user to precisely and effectively identify the nature and extent of the modifications that were made to the model and the method to consider those modifications in its estimate.

The QQR method was validated using a set of re-validation scenarios, each describing a model, the modifications made to it, and the decision to be based on the model. The QQR method's re-validation recommendations for the scenarios were compared to those of a set of human experts who were selected based on their expertise and experience in model validation. A suitable statistical measure of correlation between the QQR method's recommendations and the experts' recommendations for the scenarios was calculated. It showed strong positive correlation between the method and the experts.

This Page Intentionally Left Blank

**"Defining the Meaning of a Major Modeling and Simulation Change as Applied to Accreditation"**

## Table of Contents

## 1. Executive summary

Increasingly, decisions are being made based partially or entirely on models. These decisions may be quite important, with the potential of serious or unacceptable consequences if an incorrect decision is made. The models used to support the decisions may have been newly developed or modified versions of existing models. In the former case, it seems clear that a new model should undergo validation before it is used for any significant application. In the latter case, the matter may be less clear; must the modified model, which presumably was validated when it was initially developed, be subject to another round of validation due to the modifications?

Several existing methods address the re-validation question, including methods developed by the Johns Hopkins University Applied Physics Laboratory, the Institute of Electrical and Electronics Engineers, the Joint Accreditation Support Agency, and the project's sponsoring agency. These existing methods vary widely in several respects including level of detail, specificity with respect to modeling and simulation, degree of quantitativeness, ease of use, and applicability to the sponsoring agency. All include some form of the notion of risk, which is conventionally defined as the product of the likelihood of an incorrect decision and the consequences of such a decision.

A new method, the Quantitative-to-Qualitative Risk-based (QQR) method, was developed to make a quantitative recommendation regarding the re-validation of a modified model. The QQR method was developed with these goals in mind: to be quantitative, repeatable, and transparent; to consider both model modifications and model use risk; to focus on model types and simulation applications of interest to the sponsoring agency; and to be simple and accessible so as to encourage its use in practical applications.

The QQR method estimates the probability that not re-validating a modified model will lead to unacceptable consequences, given the modifications made to it. That estimated probability is meant to be interpreted, in the context of a re-validation decision, as a quantitative recommendation to re-validate the model. The method is based on a conditional probability formula that separates the various parts of the estimated probability into distinct terms and factors, and it provides procedures for estimating each term and factor. A central feature of the QQR method is a missions-means decomposition that allows both the method's user to precisely and effectively identify the nature and extent of the modifications that were made to the model and the method to consider those modifications in its estimate.

The QQR method was validated using a set of re-validation scenarios, each describing a model, the modifications made to it, and the decision to be based on the model. The QQR method's re-validation recommendations for the scenarios were compared to those of a set of human experts who were selected based on their expertise and experience in model validation. A suitable statistical measure of correlation between the QQR method's recommendations and the experts' recommendations for the scenarios was calculated. It showed strong positive correlation between the method and the experts.

# 2. Introduction

This section introduces this report. It states the report's purpose, identifies the authors, acknowledges other contributors, describes the report's structure and content, and makes a disclaimer regarding the report's content. It also provides a summary timeline of the project.

## 2.1 Report purpose

This report was prepared for the U. S. Department of Defense by the University of Alabama in Huntsville (UAHuntsville) Center for Modeling, Simulation, and Analysis (CMSA). It is the deliverable final report for the Systems Engineering Research Center Research Task 38, titled "Defining the Meaning of a Major Modeling and Simulation Change as Applied to Accreditation". The project contract number is H98230-08-D-0171. The UAHuntsville organization number for that project is 675250 and its fund/grant number is 26991.

## 2.2 Report authors

This report was written by (in alphabetical order):
Philip W. Alldredge, University of Alabama in Huntsville
J. Cameron Beach, University of Alabama in Huntsville
Wesley N. Colley, Ph.D., University of Alabama in Huntsville
Mikel D. Petty, Ph.D. , University of Alabama in Huntsville

## 2.3 Acknowledgements

Funding for this project was provided by:
An agency of the U. S. Department of Defense

This project received administrative support from:
Doris H. Schultz, Systems Engineering Research Center, Stevens Institute of Technology

Additional technical contributions to the project were are made by (in alphabetical order):
Employees of an agency of the U. S. Department of Defense
Juliana L. Fortune, Ph.D., University of Alabama in Huntsville[1]
Joseph G. Kovalchik, Ph.D., Johns Hopkins University Applied Physics Laboratory
Antonio McInnes, University of Alabama in Huntsville[2]
Peter P. Pandolfini, Ph.D., Johns Hopkins University Applied Physics Laboratory
Gregory S. Reed, University of Alabama in Huntsville
James J. Swain, Ph.D., University of Alabama in Huntsville

## 2.4 Report content disclaimer

The entire content of this report, including all assessments and findings stated within it, is the product and responsibility solely of its authors. No endorsement of or agreement with the content of this report by the Systems Engineering Research Center, the Stevens Institute of Technology, the technical contributors to the project acknowledged earlier, the project's sponsoring agency, the employees of the sponsoring agency, or the U. S. Department of Defense is intended or should be inferred.

[1] Dr. Fortune has since left UAHuntsville.
[2] Mr. McInnes has since left UAHuntsville.

## 2.5 Project timeline

- Project period of performance begins 2-24-2012
- J. Fortune, initial Principal Investigator, leaves UAHuntsville and project 5-11-2012
- M. Petty begins working as interim Principal Investigator 5-14-2012
- M. Petty submitted for formal approval as new Principal Investigator 5-14-2012
- Progress and status review (teleconference) 5-30-2012
- A. McInnes leaves UAHuntsville and project 6-22-2012
- Progress and status review (teleconference) 6-22-2012
- Interim project review conducted (face-to-face meeting at UAHuntsville) 6-26-2012
- M. Petty formally approved as new Principal Investigator 7-24-2012
- Progress and status review (teleconference) 8-8-2012
- Progress and status review (teleconference) 8-21-2012
- Progress and status review (teleconference) 9-19-2012
- Project period of performance extended to 10-31-2012 9-28-2012
- Initial version of final project review presentation submitted 10-22-2012
- Final version of final project review presentation submitted 10-30-2012
- Final project review conducted (teleconference) 10-31-2012
- Project period of performance ends 10-31-2012
- Final report submitted 12-12-2012

## 2.6 Report structure and content

Following the executive summary (Section 1) and this introductory section (Section 2), Section 3 states the project's motivation and objectives. Section 4 briefly summarizes some background concepts important to the context of this report, and then briefly surveys three existing methods for determining whether and to what extent a model should undergo verification, validation, and accreditation. Section 5 details the new method, the Qualitative-to-Quantitative Risk-based (QQR) method, developed for this project to address the same question. Section 6 reports the validation performed to test the new QQR method, including both the validation process and its results. Finally, Section 7 presents the project's findings and identifies potential future work.

This report also includes a list of references, the authors' biographies, and as an optional appendix, the complete text of the validation scenarios.

# 3. Project motivation and objectives

This section explains the project's motivation and states its research objectives.

## 3.1 Project motivation

Motivated by both the increasing sophistication and accuracy of simulation models, and the increasing complexity and cost of physical testing with real systems, a growing number of decisions are being made based partially or entirely on the results of simulations [Reynolds, 2009]. Selected examples include predicting whether a proposed highway expansion will solve congestion problems, optimizing assignment of production tasks to facilities, and training physicians to diagnose specific medical conditions [Fontaine, 2009]; acquisition of defense systems [Balci, 2000]; estimating the outcome of a proposed military or economic intervention in a foreign nation [Kott, 2010]; and design of medical treatment facilities [Chetouane, 2012].

These simulation-based or simulation-informed decisions may be quite important, with the potential of serious or unacceptable consequences if an incorrect decision is made. Unfortunately, the use of a model in support of a decision is no guarantee of correctness. Two examples will suffice to illustrate the point. First, the 2008 financial crisis in the United States was precipitated in part because of a large number ill-advised investment decisions that were made based on a financial model. The model, known as the Gaussian copula, assumed that the price of a credit default swap was correlated with and could predict the price of mortgage backed securities, and was widely used by investors, issuers, and rating agencies. It ultimately proved to be invalid, with enormous financial consequences [Salmon, 2009].[3] Second, the 2003 loss of the space shuttle *Columbia* and its crew stemmed in part from the decision to set aside the results of two models developed to predict the depth of exterior panel penetration that could be caused by debris strikes during launch. The models, which were used while the *Columbia* was in orbit and correctly predicted panel penetration, had been given input debris size values outside the bounds of their previous validation and their results were consequently disregarded [Gehman, 2003].[4]

Simulation projects may either develop and test new model(s) or modify existing model(s). In the former case, it seems clear that a new model should undergo verification, validation, and accreditation (VV&A) before it is used for any serious application. In the latter case, the matter may be less clear; must the modified model, which presumably underwent VV&A when it was initially developed, be subject to another round of VV&A due to the modifications? The question arises because VV&A can be difficult and costly [Youngblood, 2000] [Hartley, 2010], and project resources may already have been expended on the effort required to develop or modify the model [Balci, 1996] [Petty, 2010].

The extent of the modification should be considered when assessing the need for re-validation. On the one hand, if the model's architecture, logic, or mathematical basis was extensively modified, then a new round of VV&A is almost certainly in order. On the other hand, if the only modification was the correction of a spelling error in an output report heading, then re-validation of the model is most likely unnecessary.

[3] This is an example of a Type II error: use of an invalid model. VV&A error types will be defined in more detail later.

[4] This is an example of a Type I error: non-use of a valid model. VV&A error types will be defined in more detail later.

The risk associated with the use of the model should also be considered. If the model will be used to support a decision that could lead to unacceptable consequences if made incorrectly, then re-validation of the model may be essential.[5] In contrast, if the model will be used only in support of a real-world decision that would have negligible consequences if made incorrectly, then an expensive re-validation process may be unnecessary.

The preceding discussion leads to the project's motivating questions:

1. When should a modified model undergo re-validation?
2. How extensive must the modifications to the model be, or how serious must the consequences of the use of an incorrect model be, before the time and expense of a re-validation of the model is justified?
3. Is it possible to quantify the considerations of model modification extent and model use risk and use those quantifications to produce a quantitative recommendation for model re-validation?

## 3.2 Research objectives

This project has four specific research objectives:

1. Identify, analyze, assess, and synthesize past work relevant to the question of whether a modified model should undergo a new VV&A process.
2. Develop a new quantitative, repeatable, and transparent method to make a quantitative recommendation for the re-validation decision. The new method should consider both model modification extent and model use risk, and it should focus on model types and simulation applications of interest to the sponsoring agency. The new method should be simple and accessible to encourage use in practical applications.
3. Implement a proof-of-principle prototype of software supporting the new method.
4. Test and validate the new method by using it to make re-validation recommendations in a controlled experimental setting.

[5] For clarity, it should be mentioned explicitly that there are two decisions being discussed here. One is the decision whether or not to re-validate the model. The other is the decision regarding a real-world system that the model is being used to inform. Here the decision that is meant will be identified explicitly (as the "re-validation decision" or the "real-world decision") when it is not implicitly clear from context.

## 4. Background concepts and existing methods

This section begins by briefly reviewing three background concepts important to this report and the research it describes: the definitions of verification, validation, and accreditation; the types of VV&A errors; and the concept of risk in a VV&A context. This section then surveys three existing methods for determining whether and to what extent VV&A may be required.[6]

Two of the three existing methods (MURM and JASA) are specific to modeling and simulation (M&S), i.e., they focus on VV&A of simulation models. The third existing method (IEEE P1012) is more general, in that it addresses VV&A and software and hardware systems in general. All three of the existing methods have useful ideas and features which were ultimately included in the new method developed in this project.

### 4.1 Background concepts[7]

In general terms, *verification* refers to a testing process that determines whether a product is consistent with its specifications or compliant with applicable regulations. In modeling and simulation, verification is typically defined analogously, as the process of determining if an implemented model (and its associated data) is consistent with its conceptual description and specification [DOD, 2009]. Verification is also concerned with whether the model as designed will satisfy the requirements of the intended application. Verification examines transformational accuracy, i.e., the accuracy of transforming the model's requirements into a conceptual model and the conceptual model into an executable model. The verification process is frequently quite similar to that employed in general software engineering, with the modeling aspects of the software entering verification by virtue of their inclusion in the model's design specification. Typical questions to be answered during verification include:

1. Does the program code of the executable model correctly implement the conceptual model?
2. Does the conceptual model satisfy the intended uses of the model?
3. Does the executable model produce results when needed and in the required format?

In general terms, *validation* refers to a testing process that determines whether a product satisfies the requirements of its intended customer or user. In modeling and simulation, validation is the process of determining the degree to which the model (and its associated data) is an accurate representation of the simuland, with respect to the intended uses of the model [DOD, 2009].[8] Validation examines representational accuracy, i.e., the accuracy of representing the simuland in the conceptual model and in the results produced by the executable model. The process of

[6] The brief descriptions of the existing methods given in this report are meant only to provide a basic familiarity with those methods. Due to their brevity, these descriptions can not convey the full scope and capabilities of the methods. Readers interested in the existing methods should not consider these descriptions to be definitive and are encouraged to consult the source documents cited for each method.

[7] The text of this section is adapted and extended from [Petty, 2010].

[8] A *simuland* is the real-world item of interest. It is the object, process, or phenomenon to be simulated. The simuland might be the aircraft in a flight simulator (an object), the assembly of automobiles in a factory assembly line simulation (a process), or underground water flow in a hydrology simulation (a phenomenon). The simuland may be understood to include not only the specific object of interest, but also any other aspects of the real-world that affect the object of interest in a significant way. Simulands need not actually exist in the real world; for example, in combat simulation hypothetical non-existent weapons systems are often modeled to analyze how a postulated capability would affect battlefield outcomes.

validation assesses the accuracy of the models.[9] The accuracy needed should be considered with respect to its intended uses, and differing degrees of required accuracy may be reflected in the methods used for validation.[10] Typical questions to be answered during validation include:

1. Is the conceptual model a correct representation of the simuland?
2. How close are the results produced by the executable model to the behavior of the simuland?
3. Under what range of inputs are the model's results credible and useful?

*Accreditation*, although often grouped with verification and validation in the modeling and simulation context in the common phrase "verification, validation, and accreditation", is an entirely different sort of process from the others. Verification and validation are fundamentally testing processes, and are technical in nature. Accreditation, on the other hand, is a decision process, and is non-technical in nature, though it may be informed by technical data. Accreditation is the official certification by a responsible authority that a model is acceptable for use for a specific purpose [DOD, 2009]. Accreditation is concerned with official usability, i.e., the determination that the model may be used. Accreditation is always for a specific purpose, such as a particular training exercise or analysis experiment, or a particular category of applications. Models should not be accredited for "any purpose" because an overly broad accreditation could result in a use of a model for an application for which it has not been validated or is not suited. The accrediting authority typically makes the accreditation decision based on the findings of the verification and validation processes. Typical questions to be answered during accreditation include:

1. Are the capabilities of the model and requirements of the planned application consistent?
2. Do the verification and validation results show that the model will produce usefully accurate results if used for the planned application?
3. What are the consequences if an insufficiently accurate model is used for the planned application?

To summarize these definitions, note that verification and validation are both testing processes, but they have different purposes.[11] The difference between them is often summarized in this way: verification asks "Was the model built right?", whereas validation asks "Was the right model built?" [Balci, 1998a] [Balci, 1998b]. Continuing this theme, accreditation asks "Is the model appropriate for a particular use?" [Youngblood, 2000].

[9] Validation is used to mean assessing a model's utility with respect to a purpose, rather than its accuracy with respect to a simuland, in [Cohn, 2009] (pp. 200-201). That meaning, which has merit in a training context, is not used here.

[10] The validity of a model is always with respect to its intended use. The same model may be valid (or sufficiently valid) for one use and invalid (or insufficiently valid) for another use. Hereinafter, whenever a model's validity is discussed, the reference to validity should be understood as with respect to intended use, even if the intended use is not mentioned explicitly.

[11] Verification and validation are concerned with accuracy (transformational and representational, respectively), which is only one of several aspects of quality in a simulation project; others include execution efficiency, maintainability, portability, reusability, and usability (user friendliness) [Balci, 1998a].

| | Model valid | Model not valid | Model not relevant |
|---|---|---|---|
| Model used | Correct | **Type II error**<br>Use of invalid model;<br>Incorrect V&V;<br>Model user's risk;<br>**More** serious error | **Type III error**<br>Use of irrelevant model;<br>Accreditation mistake;<br>Accreditor's risk;<br>**More** serious error |
| Model not used | **Type I error**<br>Non-use of valid model;<br>Insufficient V&V;<br>Model builder's risk;<br>**Less** serious error | Correct | Correct |

*Table 1. Types of VV&A errors* [Balci, 1981] [Balci, 1985] [Balci, 1990] [Balci, 1998a].

Verification and validation are non-trivial processes, and there is the possibility that they may not be done correctly in every situation. What types of errors may occur during verification and validation, and what risks follow from those errors? Table 1 summarizes the types of verification and validation errors and risks.[12] In the table, three possibilities regarding the model's accuracy are considered; it may be accurate enough to be used for the intended application ("valid"), it may not be accurate enough ("not valid") or it may be not relevant to the intended application. Two possibilities regarding the model's use are considered; the model's results may be accepted and used for the intended application, or they may not. The correct decisions are, of course, when a valid model is used or when an invalid or irrelevant model is not used.

A Type I error occurs when a valid model is not used. For example, a valid flight simulator is not used to train and qualify a pilot. This may be due to insufficient validation to persuade the accrediting authority to certify the model for that application. A Type I error can result in model development costs that are entirely wasted if the model is never used or needlessly increased if model development continues [Balci, 1998a]. Additionally, whatever potential benefits that using the model might have conferred, such as reduced training costs or improved decision analyses, are delayed or lost. The likelihood of a Type I error is termed *model builder's risk* [Balci, 1981] [Balci, 1990].

A Type II error occurs when an invalid model is used. For example, an invalid flight simulator is used to train and qualify a pilot. This may occur when validation is done incorrectly but convincingly, erroneously persuading the accrediting authority to certify the model for use. A Type II error can result in disastrous consequences, such as an aircraft crash because of an improperly trained pilot or a bridge collapsing because of faulty analyses of structural loads and stresses. The likelihood of a Type II error is termed *model user's risk* [Balci, 1981] [Balci, 1990].

[12] The figure is adapted from a flowchart that shows how the errors might arise found in [Balci, 1998a]. A similar table appears in [Banks, 2005].

A Type III error occurs when an irrelevant model, i.e., one not appropriate for the intended application, is used. This differs from a Type II error, where the model is relevant but invalid; in a Type III error the model is, in fact, valid for some purpose or simuland, but it is not suitable for the intended application. For example, a pilot may be trained and qualified for an aircraft type in a flight simulator valid for some other type. Type III errors are distressingly common; models that are successfully used for their original applications often acquire an unjustified reputation for broad validity, tempting project managers, eager to reduce costs by leveraging past investments, to use the models inappropriately. Unfortunately, the potential consequences of a Type III error are similar, and thus similarly serious, to those of a Type II error. The likelihood of a Type III error is termed *model accreditor's risk* [Balci, 1990].[13]

The notion of *risk* is central to the existing methods to be described and to the method to be developed. It is easy to define risk in an intuitive and non-quantitative way, both in general and in the context of model VV&A. In general, risk is conventionally defined in this way:

*General risk* = (*likelihood of error*) · (*consequences of error*)

Reasonably enough, risk is a conceptual product of the likelihood of making an error and the magnitude of the consequences the error. For example, buying a lottery ticket has a high probability of error (i.e., choosing the wrong numbers) but low consequences of making the error (i.e., losing a dollar); thus the overall risk is low. In contrast, heart transplant surgery has, arguably, a moderate risk of error (e.g., surgical complications, infection, organ rejection) and high consequences of the error (e.g., death), leading to a higher overall risk than buying a lottery ticket. Finally, as an example of a problematic case in such calculations, the likelihood of a large asteroid colliding with the Earth is extremely low, but the consequences could be catastrophic; the overall risk here could be assessed as low or high, depending on the order of magnitude of the other quantities.

The general notion of risk can be extended to the modeling and simulation context and applied to the VV&A of a model in a straightforward way:

*VV&A risk* = (*probability the model is invalid*) · (*consequences of using an invalid model*)

Much like general risk, VV&A risk is a product of the probability that the model is invalid and the magnitude of the consequences of using the invalid model. For example, assume a particular flight simulator has significantly invalid flight dynamics, giving the user an incorrect understanding of how to fly a certain aircraft type; in other words the probability the model is invalid is high. If the flight simulator in question is to be used for entertainment, the risk is low, because the consequences of using it are low; if the flight simulator is to be used to train commercial pilots to fly that aircraft type, the consequences could be much higher and consequently the resulting risk is also higher.[14]

This notion of VV&A risk raises two questions that need to be answered in any complete methodology for quantifying VV&A risk:

---

[13] Some experts consider a Type III error to be a special case of a Type II error and argue against listing Type III separately (e.g., [Tolk, 2012]), but here we follow [Balci, 1990] and list it, an inclusion that seems justified based on the frequency with which this type of error occurs in practice.

[14] A zero-flight time flight simulator is one certified by the Federal Aviation Administration as suitable for training commercial pilots to fly a specific aircraft type with no actual flight time in that type required [Ford, 1997] [Kesserwan, 1999]. The first time a pilot flies an actual aircraft of the type he or she trained for in the zero-flight time simulator, there may be passengers in the cabin.

1. Should the probability that the model is invalid be input to a assessment of VV&A, calculated as part of an assessment, output from an assessment, or simply ignored and assumed to be 1.0?
2. How are the consequences of using an invalid model measured, and what units (e.g., lives, dollars, person-months) should be used to quantify those consequences?

### 4.2 MURM

The Modeling and Simulation Use Risk Methodology (MURM) was developed at the Johns Hopkins University Applied Physics Laboratory [JHU APL, 2011]. It is a methodology for scoping the effort and selecting the activities devoted to V&V in an M&S project based on the risk of making an incorrect real-world decision due to an invalid model. Intuitively, the MURM guides to the user to perform sufficient V&V to reduce the M&S use risk to an acceptable level. Therefore, the MURM combines both M&S risk assessment and VV&A activity planning in an integrated methodology.

A central concept of the MURM is M&S use risk, which is defined in the MURM as "The probability that inappropriate application of M&S results for the intended use will produce unacceptable consequences to the decision maker" [JHU APL, 2011]. As would be expected, higher M&S use risk suggests more effort devoted to V&V activities.

The MURM has two major features. The first is a series of formulas and tables that allow the user to calculate a numerical value for M&S use risk, and the second is a procedure that uses the calculated risk value to customize the V&V activities. Regarding the first of these major features, the calculation of M&S use risk, the primary formula has both logical and algebraic forms:[15]

Logical: $p[(C \wedge E) \wedge (C \Rightarrow E)]$

Algebraic: $p(C) \cdot p(E) \cdot [1 - p(C) + p(C) \cdot p(E)]$

In both forms, $C$ refers to "causes", i.e., inappropriate application of the results of an invalid model, and $E$ refers to "effects", unacceptable consequences resulting from the inappropriate application. The logical form of the formula is intended to convey that in the M&S use risk scenario being quantified, both causes and effects must be present and the causes must lead to the effects. The algebraic form is more directly usable to calculate a numerical value for M&S use risk. There $p(C)$, the probability of the causes occurring, is defined as:

$$p(C) = p(C_1 \cup C_2 \cup C_3)$$

where $C_1$ is a probability that the lack of clarity of intended use leads to misuse of the model's results, $C_2$ is the probability of an impact on the real-world decision if a model capability is not achieved, and $C_3$ is the probability of incorrect recommendation to use results. Numerical values for $C_1$, $C_2$, and $C_3$ are found using detailed qualitative-to-quantitative tables given in the MURM document. Table 2 is an example of a MURM qualitative-to-quantitative table (referred to as a *state table* in the MURM), taken directly from the MURM document. It illustrates how a set of qualitative descriptors, logically arranged, can be converted into a quantitative factor.

[15] As this is written (12-11-2012), there are several questions regarding the details of the mathematics in the MURM that are the subject of ongoing discussions between the authors of the MURM and the authors of this report. In this section the mathematics are presented as asserted in the MURM.

| Factor Level | Consequence / Mitigation | Level Weighting | $p(C_2)$ |
|---|---|---|---|
| A | Negligible consequence / Mitigation not required | 1 | 0.038 |
| B | Negligible consequence / Mitigation complete | 3 | 0.115 |
| C | Negligible consequence / Mitigation partial<br>Minor consequence / Mitigation complete | 6 | 0.231 |
| D | Negligible consequence / Mitigation impossible<br>Minor consequence / Mitigation partial<br>Serious consequence / Mitigation complete | 11 | 0.423 |
| E | Minor consequence / Mitigation impossible<br>Serious consequence / Mitigation partial<br>Grave consequence / Mitigation complete | 17 | 0.654 |
| F | Serious consequence / Mitigation impossible<br>Grave consequence / Mitigation partial | 22 | 0.846 |
| G | Grave consequence / Mitigation impossible | 25 | 0.962 |

*Table 2. Example MURM qualitative-to-quantitative table: Importance $p(C_2)$* [JHU APL, 2011].

| Factor Level | Reliance / Impact | Level Weighting | $p(E)$ |
|---|---|---|---|
| A | Supplemental use / Single low risk area | 1 | 0.025 |
| B | Supplemental use / Single medium risk area<br>Secondary use / Single low risk area | 4 | 0.100 |
| C | Supplemental use / Multiple medium-low risk area<br>Secondary use / Single medium risk area<br>Primary use / Single low risk area | 9 | 0.225 |
| D | Supplemental use / Single high risk area<br>Secondary use / Multiple medium-low risk area<br>Primary use / Single medium risk area<br>Only use / Single low risk area | 16 | 0.400 |
| E | Supplemental use / Multiple high risk area<br>Secondary use / Multiple high risk area<br>Primary use / Multiple medium-low risk area<br>Only use / Single medium risk area | 24 | 0.600 |
| F | Secondary use / Multiple high risk area<br>Primary use / Single high risk area<br>Only use / Multiple medium-low risk area | 31 | 0.775 |
| G | Primary use / Multiple high risk area<br>Only use / Single high risk area | 36 | 0.900 |
| H | Only use / Multiple high risk area | 39 | 0.975 |

*Table 3. Example MURM qualitative-to-quantitative table: Effects p(E)* [JHU APL, 2011].

In the algebraic form of the formula, $p(E)$, the probability of unacceptable consequences occurring, is found using a qualitative-to-quantitative table that considers both impact, the extent of the model results being used in the decision, and reliance, the dependence of the decision making process on the model results (as opposed to other considerations). An example of a MURM table for $p(E)$ follows as Table 3.

As mentioned, the second of the major features of the MURM is a procedure that links M&S use risk to V&V activities. The procedure is intended to calculate the M&S use risk for a given V&V plan, and provide guidance for how to plan V&V activities to reach a desired level of risk. The full details are substantial and beyond the scope of this report, but the procedure can be summarized as follows:

The MURM's M&S use risk calculation is embedded and used within a process designed to allow the scoping or tailoring of M&S activities in a way that considers the risk. The process uses a V&V plan, which is a list of selected V&V activities and the associated V&V methods. The V&V plan is an input to the calculation of M&S use risk. Within the process, the different capabilities of the model are prioritized with respect to its intended use, thereby setting requirements for the V&V plan. The steps of that process can be summarized as follows:

(1) Establish intended uses for the model.
(2) Enumerate the requirements.
(3) Prioritize requirements.
(4) Establish or select "Causes" state tables $C_1$, $C_2$, $C_3$.
(5) Establish initial V&V plan.
(6) Evaluate requirement-by-requirement state levels for $C_1$, $C_2$, $C_3$.
(7) Compute $p(C)$.
(8) Establish or select "Effects" state table.
(9) Evaluate requirement-by-requirement state levels for Effects; compute $p(E)$.
(10) Compute M&S use risk for each requirement as $f(p(C), p(E))$.
(11) Evaluate acceptability of each requirements' M&S use risk.
(12) Accept or modify V&V plan; if latter, return to step (5).

The specific details of each step in the MURM process are less important here than the fact that the process relates M&S use risk to the activities selected in the V&V plan. The V&V plan is considered in the value calculated for $p(C_3)$ in step (6) of the process and used in step (7). Associated with the V&V activities in the plan are specific V&V methods (termed "techniques" in the MURM); each method is given an effectiveness ("level" in the MURM) rating based on the M&S V&V Process Maturity Model [Harmon, 2005] (also see [Conwell, 2000]). Different V&V plans will produce different risk values; all else being equal, V&V plans with more (or less) effort devoted to V&V activities applying effective methods will have less (or more) M&S use risk. If the calculated risk value is judged to be too high in step (11), the plan is modified (e.g., additional V&V activities may be added to the plan) in step (12) before returning to step (5) to iterate.

The MURM contains a number of innovative ideas regarding the quantification of M&S use risk. It considers the risk of using an invalid model, decomposing that risk into multiple causes and effects. It does not accept as input an explicit specification or quantification of the extent or nature of the modifications made to the model; rather the MURM is intended to be applied to each capability of a model, so modifications to multiple capabilities of a model will introduce use risk for each modified capability. It considers the magnitude of the consequences of an incorrect decision in terms of their unacceptability to the decision maker.

### 4.3 IEEE P1012

The Institute of Electrical and Electronics Engineers (IEEE) P1012 Draft Standard for System and Software Verification and Validation is a draft standard is designed to lay out the criteria for conducting various V&V activities and to provide guideline for selecting specific V&V tasks within a project. The standard is not specific to M&S, and within the standard V&V are understood in a more generic sense than in the M&S context. Somewhat unusually, the standard is intended to apply to both hardware and software, or both in an integrated system. As such, it is really more of a systems engineering approach to V&V, which therefore relies more heavily on systems engineering artifacts, such as requirements documentation, than many other V&V methodologies, which typically rely on actual model outputs as compared to reality. The method interprets the V&V process to consist mainly of validating systems engineering documents, and verifying that systems are compliant with their supporting systems engineering documents. This interpretation permits (and requires) V&V over the system life-cycle.

Unlike many other V&V methodologies, this one has a particular focus on the user (of the software and/or hardware). So, rather than asking whether a system is valid in an abstract way, the methodology asks whether a system is valid from the perspective of a particular user, where the user may be defined in a number of ways.

For our purposes, the most salient aspect of this methodology is its exhaustive delineation of V&V activities and the rationale for conducting those activities (more than its philosophical interpretations of V&V within the systems engineering process).

The primary objective of the IEEE P1012 standard is to broaden the scope of V&V from (traditionally) a final step in a modeling effort up to a fully integrated systems engineering activity that is conducted throughout the full life-cycle of any system (software or hardware). This is obviously a massive expansion of the traditional scope of V&V. The necessary secondary objective is therefore to identify all the places within the life cycle that V&V should be being used, but traditionally isn't. The result is a truly exhaustive delineation of V&V activities accompanied by descriptions of when to use them—the documentation extends over 200 pages, including 46 separate tables of activities (e.g., Tables 4 and 5), before dozens more ancillary reference tables to be used in completing the main 46 tables.

The central concept in the IEEE P1012 methodol is "integrity level." This term is a roll-up of many risk-related concepts to produce a value representing project-unique characteristics, including complexity, criticality, risk, safety, security, desired performance, and reliability, that together define the importance of the system to the user. This is a broadening of the concept of risk to that of importance. The sense of "integrity" here is "how integral" the subsystem or engineering activity is to the overall system's performance, rather than the more usual "honesty" sense of the word.

The method is designed to consider all major system interfaces, and the validity thereof. These include the user/operator interface, the interface with the environment (physical or digital), and interfaces with other systems operating in the environment (such as other software programs).

The method relies on qualitative estimations of integrity level at each of the very large number of its entry points into the life cycle systems engineering process. It also recognizes the hierarchical nature of systems, but provides little guidance on how the hierarchy should frame the evaluation. The estimates are purely qualitative; there is no quantitative estimation or roll-up of estimates.

| V&V Activity IEEE 1012 | System Processes (ISO/IEC 15288-2008) | Software Processes (ISO/IEC 12207-2008) | Hardware Processes (NOTE 3) |
|---|---|---|---|
| Management of V&V | Not specified | Not specified | Not specified |
| Acquisition Support V&V | Acquisition Process | Acquisition Process | Acquisition |
| Supply Planning V&V | Supply Process | Supply Process | Supply |
| Project Planning V&V | Project Planning Process | Project Planning Process | Project Planning Process |
| Configuration Management V&V | Configuration Management Process | Software Configuration Management Process | Hardware Configuration Management |
| System Stakeholder Requirements Definition V&V | Stakeholder Requirements Definition Process | Stakeholder Requirements Definition Process | Stakeholders Requirements Definition Process |
| System Requirements Analysis V&V | System Requirements Analysis Process | System Requirements Analysis Process | System Requirements Analysis Process |
| System Architecture Design V&V<br>Software Concept V&V<br>Hardware Concept V&V | System Architectural Design Process | System Architectural Design Process | System Architectural Design Process |
| System Implementation V&V<br>Software/Hardware V&V Activities<br>— Concept V&V<br>— Requirements V&V<br>— Design V&V<br>— Implementation/Fabrication V&V<br>— Integration Test V&V<br>— Qualification Test V&V<br>— Acceptance Test V&V | System Implementation Process | Software Requirements Analysis Process<br>Software Architectural Design Process<br>Software Detailed Design Process<br>Software Construction Process<br>Software Integration Process<br>Software Qualification Testing Process | Hardware Requirements Analysis Process<br>Hardware Architectural Design Process<br>Hardware Fabrication Process<br>Hardware Test Process<br>Hardware Integration Test Process<br>Hardware Qualification Test Process |
| System Integration V&V | System Integration Process | Support System Integration Process | Support System Integration Process |
| All verification activities of the 1012 life cycle | System Verification Process | Software Verification Process | Hardware Verification Process |
| System Transition V&V<br>Software Installation and Checkout V&V (Transition)<br>Hardware Transition V&V | System Transition Process | Software Installation Process | Hardware Transition Process |
| All validation activities of the 1012 life cycle | System Validation Process | Software Validation Process | Hardware Validation Process |
| System Operation V&V<br>Software Operation V&V<br>Hardware Operation V&V | System Operation Process | Software Operation Process | Hardware Operation Process |
| System Maintenance V&V<br>Software Maintenance V&V<br>Hardware Maintenance V&V | System Maintenance Process | Software Maintenance Process | Hardware Maintenance Process |
| System Disposal V&V<br>Software Disposal V&V<br>Hardware Disposal V&V | System Disposal Process | Software Disposal Process | Hardware Disposal Process |

*Table 4. Alignment of VV&A practices to systems engineering practices in IEEE P1012.*

| V&V Activities | Activity: V&V Management (see 7.1) | | | | Activity: Acquisition Support V&V (see 7.2) | | | | Activity: Supply Planning V&V (see 7.3) | | | | Activity: Project Planning V&V (see 7.4) | | | | Activity: Configuration Management V&V (see 7.5) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Integrity Levels | Levels | | | | Levels | | | | Levels | | | | Levels | | | | Levels | | | |
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Acceptance Support | | | | | x | x | x | x | | | | | | | | | | | | |
| Configuration Management Assessment | | | | | | | | | | | | | | | | | x | x | x | x |
| Contract Verification | | | | | | | | | x | x | x | x | | | | | | | | |
| Identify Process Improvement Opportunities in the Conduct of V&V | x | x | x | x | | | | | | | | | | | | | | | | |
| Interface with other Processes | x | x | x | x | | | | | | | | | | | | | | | | |
| Management and Technical Review Support | x | x | x | x | | | | | | | | | | | | | | | | |
| Management review of the V&V effort | x | x | x | x | | | | | | | | | | | | | | | | |
| Planning the Interface between the V&V Effort and Supplier | | | | | x | x | x | x | x | x | x | x | | | | | | | | |
| Project Planning Strategy Assessment | | | | | | | | | | | | | x | x | x | x | | | | |
| Proposed/Baseline Change Assessment | x | x | x | x | | | | | | | | | | | | | | | | |
| Scoping the V&V Effort | | | | | x | x | x | x | | | | | | | | | | | | |
| System Requirements Review | | | | | x | x | x | x | | | | | | | | | | | | |
| V&V Final Report Generation | x | x | x | x | | | | | | | | | | | | | | | | |
| VVP Generation | x | x | x | x | | | | | | | | | | | | | | | | |

*Table 5. Example IEEE P1012 table to relating VV&A activities to system life-cycle stage.*

The obvious strength of the method is the exhaustive definitions of many distinct V&V activities, and the documentation of when they should be performed with respect to project phases and the reasons for doing so across a system's life-cycle. The greatest weakness of the IEEE P1012 is its lack of a roll-up into either an overall assessment, or overall program of activity. The standard's authors may argue that regarding V&V as a separate activity is contrary to the intent of the standard. For the purposes of this project, this characteristic is a weakness, although in the broader scheme of V&V applicability it may be perceived as a strength. A second drawback is that almost no mention is made of the V&V cost. From a practical standpoint, one has to regard the effort related to the imposing volume of tables required for execution of this method to be a drawback. The IEEE P1012 gives an encyclopedic delineation of V&V activities with tiebacks to broader systems engineering standards. This provides a base reference set that formalizes the sometimes vague terms of V&V. The method's expansion of V&V into the hardware realm is intriguing, though not relevant for this project.

### 4.4 JASA

The Joint Accreditation Support Activity (JASA) has developed a methodology to classify the risks associated with different levels of VV&A effort.[16] The method recognizes that resources are limited, and, as such, one requires a methodology to assess whether or not VV&A is recommended given the costs of the effort and risk of not conducting VV&A. Primarily, the method is motivated by a lack of practical guidance for stake-holders on the decision to initiate a VV&A effort. The primary objective is to provide stake-holders with a practical method for determining the scope of VV&A that is required by comparing program risk to VV&A cost.

[16] We could not find an actual name for the method in its documentation; herein it will simply be referred to as the JASA method.

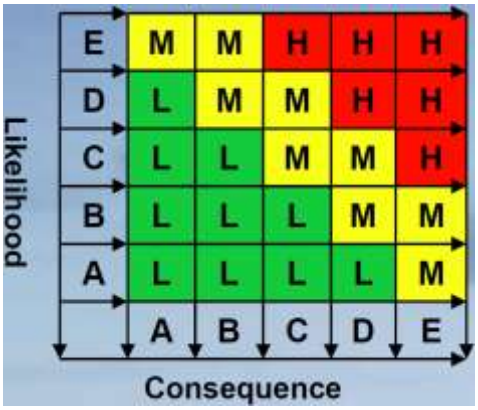


*Table 6. Risk as a qualitative "product" of likelihood and consequence in the JASA method.*

| Impact Categories | Impact Level: Catastrophic | Impact Level: Critical | Impact Level: Marginal | Impact Level: Negligible |
|---|---|---|---|---|
| Personnel Safety | Death | Severe Injury | Minor Injury | < Minor Injury |
| Equipment Safety | Major Equip Loss' Broad Scale Major Damage | Small Scale Major Damage | Broad Scale Minor Damage | Small Scale Minor Damage |
| Environmental Damage | Severe (Chernobyl) | Major (Love Canal) | Minor | Some Trivial |
| Occupational Illness | Severe & Broad | Severe or Broad | Minor and Small Scale | Minor or Small Scale |
| Cost | Loss or Program Funds; 100% Cost Growth | Funds Reduction; 50% to 100% Cost Growth | 20% to 50% Cost Growth | <20% Cost Growth |
| Schedule | Slip Reduces DoD Capabilities | Slip Causes Cost Impact | Slip Causes Internal Turmoil | Republish Schedules |
| Political | Nat'l or Internat'l (Watergate) | Significant (Tailhook) | Embarrassment ($200 Hammer) | Local |
| Operational | Widespread Add'l Combat Deaths | Limited Add'l Combat Deaths | Moderate Add'l Casualties | Minimal Add'l Casualties |

*Table 7. Consequences in the JASA method.*

The JASA method is, in principle, applicable to all software-based models; however, it appears to have been developed with defense-related models in mind.

The central concept of the JASA method is the conventional notion of risk: *risk = likelihood · consequence*. In this method, "consequence" includes the role of M&S in a decision and the real world consequence of a poor decision based on that M&S. Table 6 summarizes the JASA concept of risk as a qualitative product of likelihood of an incorrect decision based on an invalid model and of the consequences of such a decision. In the table, "A" denotes the lowest level of likelihood or consequence, and "E" denotes the highest.

This method encounters the same difficulties identified in all of the others, which is that it is nearly impossible to quantify the risk, even breaking out likelihood and consequence separately. The method relies on a set of qualitative estimates that are essentially Likert scales for qualifying the likelihood and consequences (separately). The method provides guidelines for assessing these qualitative values that are very broadly worded, but well thought-out, and, in the cases of consequences, tied back to existing military doctrine; those consequences are shown in Table 7.

The primary strengths of the method are the practicality of the technique, combined with the thoughtfulness of the tabular guidelines. The largest drawback with this method is that there is really no breakdown of individual elements and/or subsystems of models that may drive the risk—the user simply provides an overall assessment for the entire model system. The relevance of the JASA method to the current project is the strength mentioned above. We have adopted the basic approach to assessing consequences, and have adopted some of the detailed language in the consequences guidelines provided in the JASA methodology.

### 4.5 Sponsoring agency

The method currently in use by the sponsoring agency to make re-valdation decisions is highly practical. A straightforward score-sheet is used by an evaluator to assess the necessity of validation. The method differs from the others in that it specifies that verification should be conducted by the developers, rather than the stakeholders or independent evaluators; some other methods are, at best, largely silent on the specifics of *who* should conduct any particular V&V activity.

The method has a noteworthy emphasis on data validation, something that is too rarely identified specifically as a necessary component of validation. Some sources in the literature appear to regard that a model can be valid nearly independent of the input data used by the model to produce results. While this is certainly true of some simple models (e.g., a model free fall under Newtonian gravity in a vacuum), it cannot be assumed to be true in the presence of even moderately complex interactions that depend on the state of the modeled system.

The method somewhat briefly calls for "results validation," the form of validation in which model outputs are compared to the system being modeled. Often, this is the most extensive form of validation. However, for the purposes of determining the *whether* re-validation should be performed, rather than *how* it should be performed, this is not an issue.

The technique involves five scoresheets to be completed by an evaluator (an analyst or decision maker). Each of the scoresheets has several Criterion Sequence Numbers (CSNs) that identify the category in which a validity assessment is to be made. The five scoresheets are:

1. Intended Use Statement (5 CSNs). Determines whether the intended uses of the model adequately and clearly stated, both in terms of applicability and fidelity.
2. Requirements (4 CSNs).
    a Traceability: "Requirements are traceable from initial model concepts and design through testing and execution."
    b Completeness: "Requirements fully cover the functionality intended by the desired capabilities."
    c Clarity: "Requirements are unambiguous and do not require guessing or assuming information to design the functionality."
    d Consistency: There are "no conflicting requirements."
    e Testability: "It is possible from evidence generated during model development to determine explicitly if the requirement has been met."
3. Conceptual Model (8 CSNs). "The conceptual model describes how the Developer understands what is to be represented by the simulation (e.g., entities, actions, tasks, processes, and interactions) and how that representation will satisfy the requirements to which the model responds."

4. Data (12 CSNs). "This is an examination of the interaction between data and code that produces model results."
5. Results Validation (2 CSNs). "Analysis will determine if the model outputs are sufficiently accurate and realistic to meet the needs of the application."

Each CSN has a measure of performance which the evaluator must score in terms of a discrepancy; choices are none, minor, and significant. Specific types of discrepancies are delineated. Examples include "will not function," and "commits security violation" as "significant", and "documentation errors" and "could be improved" as "minor." Ultimately the set of discrepancy evaluations for all the CSNs is evaluated by an accreditation board to determine the course of action. The accreditation board will consider both the apparent necessity of re-validation as found from the CSNs, any available artifacts for prior valiation, and the estimated cost of the required revalidation effort.

The primary strengths of this method are the succinctness and usability of the evaluation. The scoresheets are straightforward to use and the report is easy to interpret. A major secondary strength, as described above, is the attention to data validity. Potential drawbacks of the method are the lack of a mathematical "roll-up" of the CSN discrepancies and a relative lack of specificity of consequences for use of an invalid model. The main relevance for this project is the guidance on scope, evaluator load, and level of detail desired by the sponsoring agency. The pragmatism, simplicity, and accessibility of this method were taken as guidelines for developing the new method.

### 4.6 Selected additional relevant literature

The literature on verification and validation has been characterized by experts in the field as "extensive" [Hartley, 2010]. A comprehensive survey of the subject is well beyond the scope of this report.[17] Here we mention selected examples from the literature that specifically address the risk of using an invalid model and/or the need to re-validate a model that has been modified.

Obviously, the documentation of the existing methods surveyed earlier consider risk and re-validation, albeit in different ways [Elele, 2007] [JHU APL, 2011] [IEEE, 2011]. As discussed earlier, three types of M&S risk (Type I, Type II, and Type III) are defined in [Balci, 1981], [Balci, 1985], [Balci, 1990], and [Balci, 1998a], and those types are discussed and applied in non-quantitative ways in [Petty, 2009], [Petty, 2010], and [Tolk, 2012]. Statistical methods to measure and control the probabilities of Type I and Type II errors are discussed in [Sargent, 2000]. [Liu, 2005] asserts that M&S use risk is reduced with more quantitative validation methods, e.g., decision risk decreases when using statistical validation as opposed to face validation. In [Chetouane, 2012], the Partitioned Multiobjective Risk Method is applied to simulation-based decision making regarding the design of a medical treatment facility. In that method, the conditional expected value of the risk associated with a typical event is complemented with the conditional expected value of the risk associated with extremely adverse events [Asbeck, 1984].

[Hartley, 2010] decomposes verification and validation risk into seven types (rather than the three defined earlier) and discusses verification and validation of models used to estimate the outcome of an international intervention, clearly an application area where the risk of using an invalid model could be quite large. In addition, [Hartley, 2010] explicitly recognizes the

[17] For useful surveys, see [Youngblood, 1993], [Balci, 1998a], [Balci, 2001], or [Petty, 2010].

modifications to or new applications of a previously validated model may necessitate re-validation (and re-accreditation) of the model.

The risk of making an incorrect and "possibly catastrophic" decision based on a model is recognized in [Balci, 2000], where approaches to planning VV&A activities intended to mitigate that risk (among other goals) are detailed, as well as in [Hale, 2007], [NASA, 2008], and [DOD, 2009]. The management and mitigation of risk was a focus of the Defense Modeling and Simulation Office's VV&A program, as described in [Youngblood, 2000]. Independent verification and validation (i.e., performed by an agent independent of the developer) is advocated as a means to reduce risk in [Balci, 2002].

# 5. Qualitative-to-Quantitative Risk-based method

This section details the new Qualitative-to-Quantitative Risk-based (QQR) method. It covers the method's design intent, the structure of its main probability calculation, the details of the terms and factors within that formula, and how the output of the formula is mapped to a qualitative re-validation recommendation. The prototype implementation of the QQR method is also described.

## 5.1 Design intent

The QQR method is intended to make a quantitative recommendation as to whether a model that has been modified should undergo re-validation. In doing so, the method will consider the changes that have been made to a model, the likelihood that the model has been rendered invalid as a result of those changes, how important the model is in making some real-world decision, and the consequences if that real-world decision is made incorrectly.

The design of the QQR method is meant to accomplish three goals:

1. Cover the entire re-validation decision, from the modifications to the model to the consequences of a real-world decision made using the model.
2. Separate the different parts of the re-validation decision so that each part and the contribution each makes to the overall re-validation decision may be considered independently.
3. Provide a mechanism to convert qualitative considerations in the re-validation decision into quantitative values.

To varying degrees, the existing methods have the same intentions, and they each have useful concepts and features relevant to the re-validation question. The development of the QQR method was intended to identify, adapt, and integrate the elements of those existing methods into a new method that would be both as simple as possible while still addressing the requirements of the sponsoring agency and remaining specific to the classes of models and types of modeling applications used by the sponsoring agency.

## 5.2 Overall formula structure

The QQR method is organized around a central formula. That formula is a conditional probability expression that both summarizes the QQR method and decomposes it into a series of six separate terms, each estimated separately. The inputs to the formula are the user's expert assessments of each term, converted from qualitative assessments to quantitative probability values by the tables and formulas of the method. The output of the formula is an estimate of the probability that not re-validating the modified model will lead to unacceptable consequences, given the modifications made to it. That estimated probability is meant to be interpreted, in the context of a re-validation decision, as a recommendation to re-validate; higher probabilities of unacceptable consequences equate to stronger recommendations to re-validate.

| QQR main formula term | MURM | JASA | IEEE |
|---|---|---|---|
| p(*modifications*) | | | |
| p(*uncertain* \| *modifications*) | | **X** | **X** |
| p(*invalid* \| *uncertain*) | **X** | | |
| p(*incorrect* \| *invalid*) | **X** | | |
| p(*unacceptable* \| *incorrect*) | | **X** | **X** |

*Table 8. Sources of the terms in the QQR main formula.*

The QQR method's "main" formula is:[18]

p(*unacceptable* | *modifications*) =
p(*modifications*) ·
p(*uncertain* | *modifications*) ·
p(*invalid* | *uncertain*) ·
p(*incorrect* | *invalid*) ·
p(*unacceptable* | *incorrect*)

As a probability expression, the QQR main formula always has a value in the interval [0, 1]. Higher values, i.e., values closer to 1, indicate that unacceptable consequences are more likely. The terms within the main formula are also probability expressions and thus they all have values in the same interval. It should be noted that each conditional probability term in the formula assumes that the previous term is true. (See Appendix A for a detailed mathematical justification of this assumption structure.) For example, p(*incorrect* | *invalid*) assumes that the model is invalid and estimates the resulting probability that an incorrect real-world decision is made as a result. The question of the truth of the assumption, i.e., whether the model is invalid, is addressed by the preceding terms.

The individual terms in the QQR main formula will each be explained in turn; p(*modifications*) in this section, p(*uncertain* | *modifications*) in the next section, and the remaining terms in the following section. The term p(*modifications*) is the easiest of all and is included so as to have a complete chain of conditions from modifications to consequences, i.e., for mathematical completeness. It denotes the probability that the model has been modified, and can have the value 0 (not modified) or 1 (modified). The QQR method is designed to be applied only if the model has been modified, so in the implementation of the QQR method it is assumed that p(*modifications*) = 1.

The existing methods each estimate, either qualitatively or quantitatively, the probabilities associated with a subset of the terms within the QQR main formula.[19] None of the existing methods appears to address all of the terms. Table 8 summarizes the relationship between the existing methods and the terms of the QQR main formula; in the table, an X indicates that the method estimates the term.

[18] The notation p(*A* | *B*) means the probability of *A*, given *B*. For introductory mathematical background on conditional probabilities, see [Gersting, 2007] or [Epp, 2011].

[19] Indeed, it was by studying those methods, identifying the parts of the re-validation decision that each considered, and then combining those parts that that QQR main formula was developed.

## 5.3 Missions-means decomposition

The term p(*uncertain* | *modifications*) denotes the probability that the model's validity with respect to its intended use is uncertain, given the modifications made to it.[20] As noted earlier, it is easy to imagine examples of modifications that introduce little or no uncertainty (e.g., correcting a spelling error in a report heading generated by the model) or that introduce full or nearly full uncertainty (e.g., radically restructuring the logic of the model and replacing the probability distributions used for all stochastic elements). These extreme examples illustrate why this term exists, but they are of little value in helping to develop a useful way of calculating a quantitative value for p(*uncertain* | *modifications*) under more typical conditions. The difficult lies in finding an unambiguous, compact, and usable mechanism for specifying the modifications that have been made to a model from among the vast number of possibilities.

To do so, the quantification of p(*uncertain* | *modifications*) in the QQR method relies on the notion of a missions-means decomposition. In other applications missions-means decompositions have been used for a range of purposes, including relating military missions and warfighting products [Sheehan, 2003] and for identifying test cases in a model verification methodology [Neal, 2005]. For example, Figure 1 shows a portion of a notional missions-means decomposition in a military setting. In the figure, missions are hierarchically decomposed on the horizontal axis, while means are hierarchically decomposed on the vertical axis. Stars represent evidence items where a particular means supports a particular mission. The missions axis has been filled out using the Uniform Joint Task List (UJTL).[21] The means axis represents the various weapon systems or other military capabilities that can be brought to bear in support of a mission. Each cell in the decomposition represents the opportunity for a particular means to support a particular mission. Some parts of the matrix may be sparse in the military context, because most missions are not supported by a particular means (e.g., tanks do not support combat air patrol missions).[22]

For the purposes of this project, our focus is not military operations but computational and network infrastructure; possible modifications to real-world network communications and computer systems are regarded as the "missions." The means axis, rather than delineating military capabilities, describes particular modeling components used to represent information technology infrastructure; more specifically, for this project the focus on an OPNET type model for the means axis.

[20] The term *uncertain* has certain technical connotations (e.g., in the field of uncertainty quantification) that are not meant here. When a model has been modified, those modifications have either rendered the model invalid, or they have not, with respect to the model's intended use. (We recognize that invalidity is not necessarily a binary state, but we make that assumption to simplify this explanation.) Therefore, it is misleading to say that there is a probability that the model in invalid as a result of the changes; it either is or it isn't, i.e., the probability is 1 or 0. However, because the model has not yet been re-validated, it is unknown whether the model is invalid. Therefore, the term uncertain measures the method's best estimate of whether the modifications have rendered the model invalid or not, expressed as a probability. The probability is estimated from the modifications via the missions–means decomposition.

[21] This source has the advantage that it not only delineates the types of missions, but also provides the relevant metrics for those missions.

[22] The authors of this report were members of a team that executed the missions-means approach successfully in testing an operational level training simulation in the context of command training at Pacific Command's Terminal Fury '05 exercise [Neal, 2005].

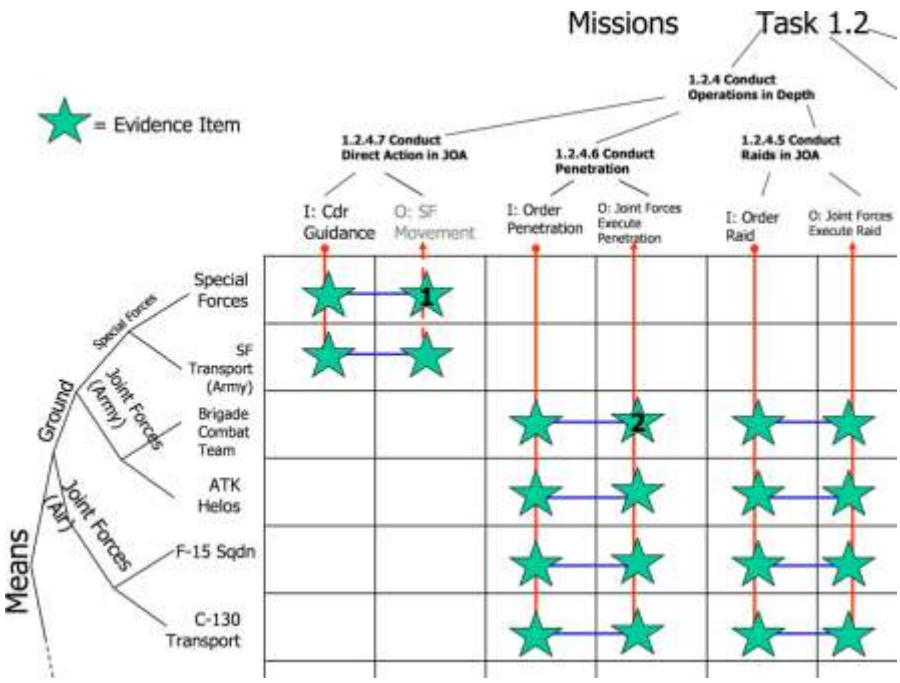


*Figure 1. Portion of a notional missions-means decomposition.*

In the QQR method, a *mission* is some specific aspect of the real world that may need to be represented in a modeling application of the types under consideration. A *mean* is a specific feature, function, capability, or parameter that could be present in a model of the class being considered. Taken together, the combination of a mission and a mean is an aspect of reality represented by a model feature; each such combination is referred to as an *evidence item*. Note that given a set of missions and a set of means, not every (mission, mean) pair that could be formed from the cross product of the sets is feasible; an evidence item exists only when it makes sense to represent the aspect of the real world corresponding to the mission with the model feature corresponding to the mean.

A missions-means decomposition is used in the QQR method in the following way. A set of missions is developed in advance. The set contains all of the missions relevant to the real world systems represented in the types of modeling applications under consideration. For this project, the means are features of networked communications and computer systems that are typically the real-world systems being modeled. A set of means is developed in advance containing all of the means relevant to the class of models under consideration. For this project, the means are features and functions available in the OPNET modeling environment.[23] From those two sets, those mission–mean pairs (i.e., evidence items) that are feasible are identified and a probability $p_i$ is associated with each evidence item $i$. Then, when the QQR method is applied to a specific situation, the evidence items that correspond to the modifications made to the model are selected by the user from among all of the evidence items available. The probabilities associated with the selected evidence items are "rolled up" to compute p(*uncertain* | *modifications*).

[23] The restriction of the missions to features of networked communications and computer systems and the restriction of the means to features and functions available in the OPNET environment were directed by the sponsoring agency.

*Figure 2. Missions-means spreadsheet.*

Figure 2 shows an example of how the mission-means decomposition is used in this project's spreadsheet implementation. Missions are on the horizontal axis and means are on the vertical axis. In the spreadsheet, the cells marked with an "X" are the possible evidence items, i.e., the feasible combinations of a mission and a mean. The cells filled with magenta are also evidence items; they were selected by the user to describe a particular set of modifications to a model (the evidence items marked with an "X" could have been selected, but were not for this example). The numeric values in the magenta cells represent the scale or magnitude of each revision, or equivalently, the degree of uncertainty with respect to model validity introduced by the particular modification represented by the evidence item, with 1 being smallest and 5 being largest.

The user is to identify evidence items in the missions-means decomposition that are relevant to changes in the model for the use in question. For each evidence item identified, the user must then evaluate on a scale of 1–5 the level to which the change is likely to introduce uncertainty. The maximum entropy conversions are then applied for all of the entered values to produce the probability values $p_i$ mentioned earlier for each modification. The implementation (which will be described in more detail later) is an Excel spreadsheet that automatically rolls up the values to produce a final output for p(*uncertain* | *modifications*).

One of the difficult aspects of the QQR method is the matter of rolling up the multiple probabilities from multiple modifications to form a final output probability for p(*uncertain* | *modifications*). Several different possible means for rolling up those probabilities were considered. The most pessimistic assumption is that the uncertainties are mutually exclusive. In this case the probabilities simply sum. This assumption quickly becomes implausible because the sum of the probabilities exceeds 1. The assumption also seems strange from the outset, because it wouldn't seem that the uncertainty for one particular modification would be mutually exclusive to another in general.

Independence of the uncertainties is a more credible assumption. Here the probabilities multiply, the same way they might for many familiar problems, such as flipping a quarter many times. The difficulty mathematically is that assuming independence generates tiny probabilities fairly quickly: $0.7^{10} = 0.03$, which means that rating ten evidence items as "likely" 4s produces a probability after roll-up of 3%. Note that the more uncertain evidence items selected, the smaller the overall probability of uncertainty one obtains, which is counterintuitive. The typical solution is to consider rolling up the opposite probability (the probability that uncertainty is *not*

introduced). At the end of the multiplication, subtract back from 1. In this case, $1 - (1 - 0.7)^{10} = 0.999994$. While an improvement, this value seems improbably close to one.

Because neither of the two usual logical choices have attractive mathematical properties, several mathematical functions that may replace the above techniques were studied. The first function considered was an exponential

$$P = 1 - \exp\left(-\sum p_i\right) \tag{1}$$

where $P$ is the output rolled up probability, and $p_i$ is the probability for a single evidence item. This function has the advantage that for low values of the sum, the output is the sum, but at higher sum values, the exponential rolls over. As it turns out, however, the roll-over is nearly identical to that seen with the independence function (see Figure 3). What is needed is a function that rolls over earlier and more slowly approaches 1. The next function considered was

$$P = \frac{2}{\pi}\tan^{-1}\left(\frac{\pi}{2}\sum p_i\right) \tag{2}$$

Again, this function is proportional to the sum for low sum values, and rolls over to a maximum value of 1 for high values. However, this function approaches 1 much more slowly than the exponential (see Figure 3). Recall that the tangent function is infinite at $\pi/2$, and as such, the arctangent of any positive number can never exceed $\pi/2$. The ultiplication by $2/\pi$ ensures that the maximum value is 1. The $\pi/2$ is inside the argument so that for small values of the sum, the output is simply the sum. The example of 10 0.7 evidence item probabilities now nets an overall 0.94 probability, which we regard to be much more plausible. We have adopted this *ad hoc* technique based on its more plausible results than other techniques studied.

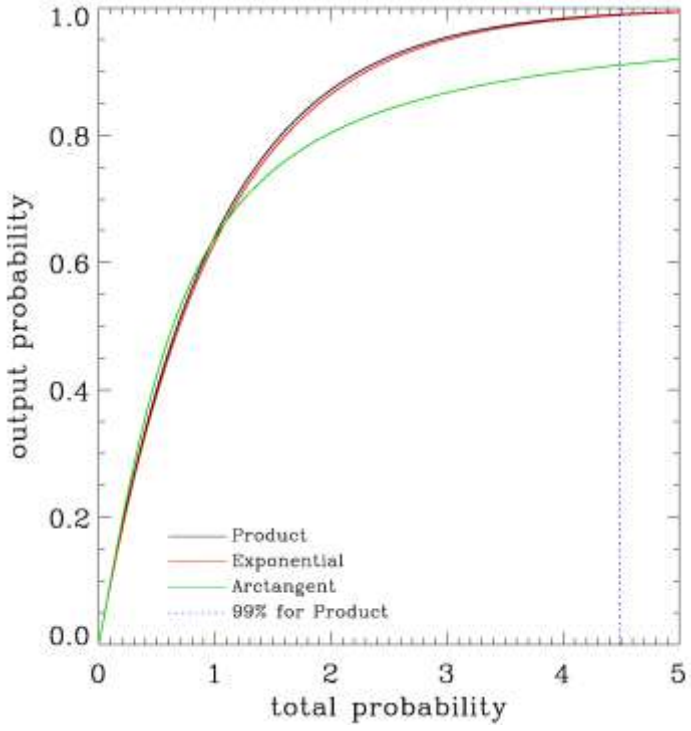


*Figure 3. Alternative probability roll-up techniques.*

## 5.4 Other terms and factors

The previous term p(*uncertain* | *modifications*) quantified as a probability the uncertainty regarding the model's validity due to the modifications made to the model. The next term p(*invalid* | *uncertain*) relates that uncertainty to the probability that the model is actually invalid with respect to its intended use.

Setting aside the constraints of practicality for a moment, it is in principle possible to construct an empirical probability distribution for this term. To do so, one could observe a large number of modeling projects, each involving modifications to a model of the class under consideration. For each project, the modifications would be noted and the missions-means decomposition would be used to calculate the uncertain value as described earlier. Then the model would be validated and its status as valid or invalid with respect to its intended application would be recorded. Once a sufficient number of projects had been observed in this manner, it is arguable possible that an empirical distribution relating the uncertainty value calculated for the modifications to the probability of discovering during validation that the model is invalid could be established using well-known input modeling methods [Banks, 2010].

Unfortunately, such a data collection effort was outside the scope of this project. Therefore, in the QQR method it will be assumed that p(*invalid* | *uncertain*) = 1. This term is included in the main formula for completeness and as a placeholder for future research.

Recall that an assumption of the QQR method is that the model being considered for re-validation may be used to support a decision regarding some real-world system, plan, or action. In other words, the results of the model may be considered by the decision maker in the process of making the real-world decision. It is possible that an invalid model could produce incorrect results that misinform the decision maker and lead him or her to make an incorrect real-world decision. The term p(*incorrect* | *invalid*) quantifies the probability that an incorrect real world decision is made, given that the model is invalid with respect to its intended use. As with the other terms in the QQR main formula, this term assumes the truth of the preceding term. Here the assumption that the model is invalid with respect to its intended use, which in this case is supporting the real-world decision. The probability the model being invalid was estimated in the previous terms.

The value for p(*incorrect* | *invalid*) is the product of four subterms, or factors: p(*applicability*), p(*reliance*), p(*credibility*), and p(*invalidity*). The formula for the term is:

p(*incorrect* | *invalid*) =
p(*applicability*) ·
p(*reliance*) ·
p(*credibility*) ·
p(*invalidity*)

| Applicability of model to decision | p(*applicability*) |
|---|---|
| Very applicable | 0.9 |
| Somewhat applicable | 0.7 |
| Neither applicable nor inapplicable | 0.5 |
| Somewhat inapplicable | 0.3 |
| Very inapplicable | 0.1 |

*Table 9. Qualitative-to-quantitative state table for QQR factor p(applicability).*

The calculation of p(*incorrect* | *invalid*) as a simple product is intentional. If any of four factors has the value 0, e.g., p(*applicability*) = 0, because the model is not at all applicable to the real-world decision being made, then the probability of making an incorrect decision because the model is invalid should be 0.[24]

The factor p(*applicability*) quantifies the degree to which the model is perceived to be applicable to the real-world decision to be made.[25] Note that higher applicability of a model to an application, such as a real-world decision, is normally positive; ideally, a model should be highly applicable to its use. However, under the assumption made in this term that the model is invalid, higher applicability is instead negative, because it contributes to the decision maker using the results of the invalid model in making the real-world decision.

The value for p(*applicability*) is found using a qualitative-to-quantitative state table, which is shown in Table 9. Note that the applicability is characterized for the user of the QQR method (in the left column) using a five-level Likert-style scale [Likert, 1932]. The choices of both the Likert-style scale and the use of five levels were intentional, as they were thought to be familiar to most potential users of the method.[26] The numerical values for p(*applicability*) for each of the levels (in the right column) were calculated using the maximum entropy method advanced and used for the same purpose in the MURM [JHU APL, 2011].

The factor p(*reliance*) quantifies the degree to which the model is relied upon or is a determining factor in the real-world decision to be made.[27] This factor recognizes and quantifies the fact that the decision maker may consider things other than the model's results (such as his or her expertise and experience, the advice of colleagues, resource constraints, and so on) in making the real-world decision.

[24] As will be seen, none of the factors of p(*incorrect* | *invalid*) can actually have the value 0, due to the maximum entropy-calculated values in the qualitative-to-quantitative state tables used to quantify the factors. The example of 0 is used here simply to illustrate the point in a clear manner.

[25] The QQR factor p(*applicability*) is related to and subsumes the MURM "clarity" factor. Counterintuitively, the perceived applicability of a model can potentially be increased by incomplete or ambiguous documentation of the intended uses of a model or its functionality, because the ambiguity may allow a user to believe the model is applicable for a use where it is not.

[26] We are aware that qualitative-to-quantitative state tables need not necessarily have five levels, and the MURM has examples of such tables [JHU APL, 2011]. We elected to consistently use five levels to make the QQR method more accessible and familiar to potential users.

[27] The QQR factor p(*importance*) is approximately equivalent to the MURM "importance" factor.

| Reliance of decision on model | p(*reliance*) |
|---|---|
| Sole determining factor | 0.9 |
| Primary determining factor among several | 0.7 |
| One of several equally important determining factors | 0.5 |
| Secondary determining factor among several | 0.3 |
| Not a determining factor | 0.1 |

*Table 10. Qualitative-to-quantitative state table for QQR factor p(reliance).*

| Credibility of model before re-validation | p(*credibility*) |
|---|---|
| Very credible | 0.9 |
| Somewhat credible | 0.7 |
| Neither credible nor not credible | 0.5 |
| Somewhat not credible | 0.3 |
| Very not credible | 0.1 |

*Table 11. Qualitative-to-quantitative state table for QQR factor p(credibility).*

The value for p(*reliance*) is found using a qualitative-to-quantitative state table, which is shown in Table 10. The table has the same structure and numerical values as the other factor tables, for the same reasons.[28]

The factor p(*credibility*) quantifies the degree to which the model is perceived to be credible, correct, or reliable even before the model is re-validated.[29] Such an assumption of model correctness may be due to trust in the model, in the modeling environment used to implement the model, or in the organization or people that developed or modified the model. To be clear, this factor does not quantify whether such trust is warranted; recall that the assumption in this term is that the model is invalid. Rather, this factor quantifies the probability that *a priori* credibility of the model will affect the re-validation decision. Note that higher credibility of a model is normally positive; if a model is valid, trust in the model helps to avoid Type I errors. However, under the assumption made in this term that the model is invalid, higher credibility is instead negative, because it contributes to the decision maker using the results of the invalid model in making the real-world decision.

The value for p(*credibility*) is found using a qualitative-to-quantitative state table, which is shown in Table 11. The table has the same structure and numerical values as the other factor tables, for the same reasons.

The factor p(*invalidity*) recognizes the fact that invalidity of a model is not a binary choice. In qualitative terms, the results calculated by an invalid model may be "slightly off" or they may be "not even close", or perhaps somewhere in between.[30] Intuitively, the degree to which an invalid model may have some effect on the probability of making an incorrect real-world decision based

[28] The similarity in table structure for the different QQR factors is intentional; the goal is to make the method more accessible to potential users.
[29] The QQR factor p(*credibility*) is included due to guidance from the sponsoring agency. They asserted that this factor is indeed considered in re-validation decisions in practice.
[30] The factor p(*invalidity*) is distinct from the term p(*invalid* | *uncertain*), despite their similar nomenclature.

on that model. Unfortunately, the degree to which the model in invalid is unknowable without validating the model, and the QQR method is designed to assist in making the re-validation decision. By the time the p(*invalidity*) might be known, i.e., after re-validation, the QQR method is no longer applicable. Therefore, in the QQR method it will be assumed that p(*invalidity*) = 1. This term is included in the main formula for completeness and as a placeholder for future research.

The final term in the QQR main formula is p(*unacceptable* | *incorrect*). This term quantifies the probability that an incorrect real-world decision that has been made using the modified and invalid model leads to unacceptable consequences. As usual, in the formula for this term it is assumed that the incorrect real-world decision has been made, and this term focuses on the consequences of that decision. The consequences of an incorrect decision are defined as the difference in cost, in terms of time, money, intelligence, and security between an incorrect decision and a correct one. In other words, even a correct decision may incur costs of time or money or lead to losses of intelligence or security; this term is focused on the additional costs of an incorrect decision beyond those of a correct decision.

The value for p(*unacceptable* | *incorrect*) is calculated from six subterms, or factors: p(*schedule slip*), p(*comm latency*), p(*intel delay*), p(*intel loss*), p(*extra cost*), and p(*security breach*).[31] Factors within p(*unacceptable* | *incorrect*) are limited to first-order consequences of an incorrect decision. For this reason, the possible loss of human lives or physical assets are intentionally not included in the factors. The formula for the term is:

p(*unacceptable* | *incorrect*) = (1 –
(1 – p(*schedule slip*)) ·
(1 – p(*comm latency*)) ·
(1 – p(*intel delay*)) ·
(1 – p(*intel loss*)) ·
(1 – p(*extra cost*)) ·
(1 – p(*security breach*)))

Each of the individual factors will be explained later. The specific mathematical formula used, with each term given as (1 – p(*x*)), rather than a simple product of the factors, is intended to produce an "additive" or aggregating effect if several of the factors are present in the consequences.[32] For example, suppose there were only two factors, p(*x*) and p(*y*), with p(*x*) = p(*y*) = 0.7. The simple product p(*x*) · p(*y*) = 0.7 · 0.7 = 0.49, giving a value for p(*unacceptable* | *incorrect*) that is less than either of the individual factors. This is the opposite of what is intended, as multiple consequences of different types should increase p(*unacceptable* | *incorrect*), not reduce it. In the example using the formula as given, (1 – (1 – p(*x*)) · (1 – p(*y*))) = (1 – (1 – 0.7) · (1 – 0.7)) = 0.91. This value is larger than either of the individual factors, but still less than 1, which is what is intended.

[31] As with the missions-means decomposition, the factors used for p(*unacceptable* | *incorrect*) are specific to this particular project and the sponsoring agency. The factors of this term contain the most sponsoring agency-specific and least general purpose values of the QQR method. If the QQR method is to be applied to another organization, both the missions-means decomposition and the factors of p(*unacceptable* | *incorrect*) will have to be replaced with constructs specific to that customer.

[32] The intent is similar to the roll-up of the missions-means probabilities, but the mathematical method chosen was different.

| Delay in project schedule | Level | p(*schedule slip*) |
|---|---|---|
| Project cannot meet key project milestones, or critical path slip ≥ 91 days | Catastrophic | 0.9 |
| Major effect on project schedule, or critical path slip ≥ 31 days and ≤ 90 days | Major | 0.7 |
| Moderate effect on project schedule, or critical path slip ≥ 16 days and ≤ 30 days | Moderate | 0.5 |
| Minor effect on project schedule, or critical path slip ≥ 5 days and ≤ 15 days | Minor | 0.3 |
| Negligible or no effect on project schedule, or critical path slip ≤ 4 days | Negligible | 0.1 |

*Table 12. Qualitative-to-quantitative state table for QQR factor p(schedule slip).*[33]

| Increase in network latency | Level | p(*comm latency*) |
|---|---|---|
| Catastrophic delay to network communications | Catastrophic | 0.9 |
| Major delay to network communications | Major | 0.7 |
| Moderate delay to network communications | Moderate | 0.5 |
| Minor delay to network communications | Minor | 0.3 |
| Negligible delay to network communications | Negligible | 0.1 |

*Table 13. Qualitative-to-quantitative state table for QQR factor p(comm latency).*[34]

The factor p(*schedule slip*) quantifies the delay in project schedule that results from an incorrect decision. The project in question is the project which is the context for or will be affected by the real-world decision; it is not the model development project.[35]

The value for p(*schedule slip*) is found using a qualitative-to-quantitative state table, which is shown in Table 12. The table has the same structure and numerical values as the other factor tables, for the same reasons.The factor p(*comm latency*) quantifies the increase in communications network latency that results from an incorrect decision. This factor makes sense in the context of the types of modeling applications under consideration, which are assumed to involve modeling a networked communications or computer system. An incorrect decision in such an application could lead to a real-world network configuration with increased latency over a similar network configured based on a correct decision.

[33] The specific values in the table (i.e., days of delay) for each level were provided by the sponsoring agency. The definitions of each level have two parts, e.g., "Major effect on project schedule, or critical path slip ≥ 31 and ≤ 90 days". The two parts are meant to be two ways of defining the same delay situation, not as two independent criteria that could have other combinations than those given in the table.

[34] Unlike the p(*schedule slip*) factor, specific values for the p(*comm latency*) factor were not provided by the sponsoring agency. Doing so may have implicitly characterized the performance or performance sensitivity of the real-world communications networks that are used by the sponsoring agency.

[35] The issue of whether the project in question was the model development project or the project which would be affected by an incorrect decision arose and was discussed during this research. To clarify, recall that at this point in the formula it is assumed that the invalid model was in fact used to make an incorrect real-world decision. An incorrect real-world decision made using a model won't affect the schedule of the model development; presumably, by the time the model is used to support a decision, its development is complete. Thus the schedule slip is with respect to the project requiring the decision, not the model development project.

| Delay in intelligence processing time | Level | p(*intel delay*) |
|---|---|---|
| Catastrophic delay in intelligence processing | Catastrophic | 0.9 |
| Major delay in intelligence processing | Major | 0.7 |
| Moderate delay in intelligence processing | Moderate | 0.5 |
| Minor delay in intelligence processing | Minor | 0.3 |
| Negligible delay in intelligence processing | Negligible | 0.1 |

*Table 14. Qualitative-to-quantitative state table for QQR factor p(intel delay).*[36]

| Cost of intelligence information loss | Level | p(*intel loss*) |
|---|---|---|
| High importance and extensive scope | Catastrophic | 0.9 |
| High importance or extensive scope | Major | 0.7 |
| Medium importance and limited scope | Moderate | 0.5 |
| Low importance and extensive scope | Minor | 0.3 |
| Low importance and limited scope | Negligible | 0.1 |

*Table 15. Qualitative-to-quantitative state table for QQR factor p(intel loss).*[37]

The value for p(*comm latency*) is found using a qualitative-to-quantitative state table, which is shown in Table 13. The table has the same structure and numerical values as the other factor tables, for the same reasons.

The factor p(*intel delay*) quantifies the increase in end-to-end intelligence processing time that results from an incorrect decision. This factor makes senses in the context of the types of modeling applications under consideration, which are assumed to be concerned with intelligence collection and analysis. An incorrect decision in such an application could lead to a real-world intelligence process that requires more time than a process based on a correct decision.

The value for p(*intel delay*) is found using a qualitative-to-quantitative state table, which is shown in Table 14. The table has the same structure and numerical values as the other factor tables, for the same reasons.

The factor p(*intel loss*) quantifies the risk or value of intelligence information loss that results from an incorrect decision. This factor makes senses in the context of the types of modeling applications under consideration, which are assumed to be concerned with intelligence collection and analysis. An incorrect decision in such an application could lead to a loss of real-world intelligence due to shortcomings of the communications network or intelligence process that would not have been lost in a network or process configured based on a correct decision.

[36] Unlike the p(*schedule slip*) factor, specific values for the p(*intel delay*) factor were not provided by the sponsoring agency. Doing so may have implicitly characterized the time durations or time sensitivity of the real-world intelligence processes that are used by the sponsoring agency.

[37] The definitions for each level depend on two considerations, importance and scope. Importance refers to the significance to national security of the lost information. Possible qualitative values for importance are high, medium, and low. Scope refers to the magnitude or extent of the information lost. Possible qualitative values for scope are extensive and limited.

| Extra project cost | Level | p(*extra cost*) |
|---|---|---|
| ≥ 100% of project cost | Catastrophic | 0.9 |
| ≥ 51% and ≤ 99% of project cost | Major | 0.7 |
| ≥ 21% and ≤ 50% of project cost | Moderate | 0.5 |
| ≥ 2% and ≤ 20% of project cost | Minor | 0.3 |
| ≤ 1% of project cost | Negligible | 0.1 |

*Table 16. Qualitative-to-quantitative state table for QQR factor p(extra cost).*[38]

| Risk and extent of security breach | Level | p(*security breach*) |
|---|---|---|
| High likelihood and extensive exposure | Catastrophic | 0.9 |
| High likelihood or extensive exposure | Major | 0.7 |
| Medium likelihood and limited exposure | Moderate | 0.5 |
| Low likelihood and extensive exposure | Minor | 0.3 |
| Low likelihood and limited exposure | Negligible | 0.1 |

*Table 17. Qualitative-to-quantitative state table for QQR factor p(security breach).*[39]

The value for p(*intel loss*) is found using a qualitative-to-quantitative state table, which is shown in Table 15. The table has the same structure and numerical values as the other factor tables, for the same reasons.

The factor p(*extra cost*) quantifies the extra or additional cost in the project that results from an incorrect decision. The project in question is the project which is the context for or will be affected by the real-world decision; it is not the model development project.

The value for p(*extra cost*) is found using a qualitative-to-quantitative state table, which is shown in Table 16. The table has the same structure and numerical values as the other factor tables, for the same reasons.

The factor p(*security breach*) quantifies the risk or value of a security breach that results from an incorrect decision. This factor makes senses in the context of the types of modeling applications under consideration, which are assumed to be concerned with intelligence collection and analysis. An incorrect decision in such an application could lead to an unintended release of sensitive information about the user's organization that could damage national security or the organization's future ability to execute its mission.

The value for p(*security breach*) is found using a qualitative-to-quantitative state table, which is shown in Table 17. The table has the same structure and numerical values as the other factor tables, for the same reasons.

[38] The specific values in the table (i.e., percentages of project cost) for each level were provided by the project sponsor.

[39] The definitions for each level depend on two considerations, likelihood and exposure. Likelihood refers to the probability information will be released. Possible qualitative values for likelihood are high, medium, and low. Exposure refers to the magnitude or extent of the information released. Possible qualitative values for scope are extensive and limited.

| Color category lower bound | Color category upper bound | Color category |
|---|---|---|
| ≥ 0 | < 0.35 | Green |
| ≥ 0.35 | < 0.5 | Yellow |
| ≥ 0.5 | ≤ 1 | Red |

*Table 18. Quantitative-to-qualitative mapping QQR output p(unacceptable | modifications).*

## 5.5 Output mapping

By design, the QQR terms, factors, and formulas collectively calculate a value for p(*unacceptable* | *modifications*) that will always be in the interval [0, 1]. That output value is meant to be interpreted as an estimate of the probability that unacceptable consequences will result if the model is not re-validated, given the modifications made to it. Larger values (i.e., closer to 1) for p(*unacceptable* | *modifications*) indicate a stronger recommendation to re-validate the model. To assist the method's user to rapidly understand the method's recommendation, the range of possible values for p(*unacceptable* | *modifications*) have been mapped to the familiar informal "green, yellow, red" scale. In this mapping, "green" indicates a situation where re-validation may be omitted with relatively little risk of unacceptable consequences, "red" indicates a situation where re-validation should not be omitted, and "yellow" indicates an intermediate situation. The values for the boundaries between the color categories are show in Table 18.

The specific values for the boundaries between the color categories were selected based on empirical review of the re-validation recommendations provided by the human experts in the method's validation process, to be described later. In other words, the boundary values were chosen so that the color categories would be consistent with the overall recommendations of the experts. The color category boundary values are certainly an area where further research and analysis, or at the very least a larger set of expert opinions, would be useful.

## 5.6 Implementation

A prototype implementation of a QQR tool was implemented as a spreadsheet, illustrated in Figure 4.

The spreadsheet has two main parts. The qualitative-to-quantitative terms and factors of the method, and its output value for p(*unacceptable* | *modifications*) can be seen in the smaller arrangement of cells in the upper left. The missions-means decomposition can be seen in the larger arrangement of cells in the lower right. As in the earlier example, missions are on the horizontal axis and means are on the vertical axis. In the spreadsheet, the cells marked with an "X" are the possible evidence items, i.e., the feasible combinations of a mission and a mean.

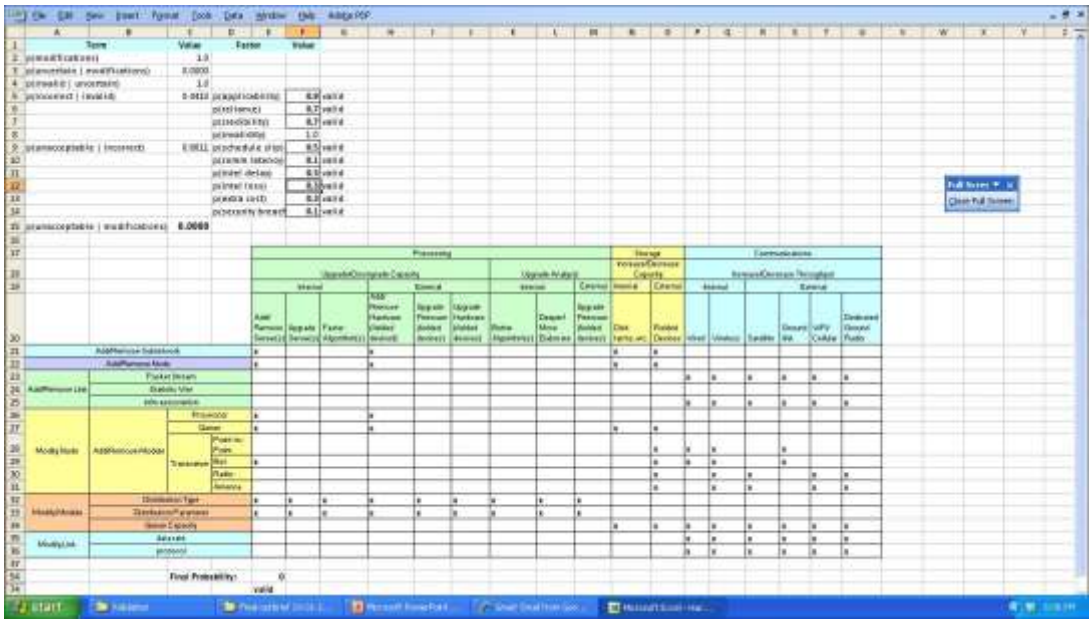

*Figure 4. Spreadsheet implementation of the QQR prototype.*

In the QQR spreadsheet prototype, non-input cells are protected and input cells are checked for valid inputs, assisting the user in using the spreadsheet properly. The user selects the appropriate evidence items for the current model modifications and enters a 1-5 Likert value for each one. The spreadsheet converts each of the user's 1–5 uncertainty assessments into a probability (.1, .3, .5, .7, .9) per the maximum entropy formula used in the MURM and rolls up the probabilities into p(*uncertain* | *modifications*) per the formula.

# 6. Validation

This section reports the validation performed to test the new QQR method, including both the validation process and its results.

## 6.1 Validation process

The QQR method's re-validation recommendations were validated to assess their accuracy.[40] Ideally, it would have been preferable to validate the QQR method's re-validation recommendations by comparing those recommendations with real-world values known to be correct. Unfortunately, such objectively correct values for re-validation recommendations were not available. As a substitute, the QQR method's recommendations were statistically compared to those of a set of humans selected based on their expertise in model validation.

The process used to experimentally test and validate the QQR method is illustrated in Figure 5. It had these steps:

(1) Develop and document a set of eight validation scenarios.
(2) Obtain recommendations from eight modeling and simulation experts regarding re-validation for each of the scenarios and variants.
(3) Quantify each of the scenarios and variants using the QQR method; for each, select the missions-means evidence items and the levels in the qualitative-to-quantitative tables for the other terms and factors.
(4) Input the quantified scenarios and variants into the QQR method spreadsheet and calculate the QQR re-validation recommendation value for each.
(5) Compare the QQR method's re-validation recommendations with those provided by the experts and calculate a statistical measure of correlation between them.

*Step (1).* Notional scenarios intended to be representative of the classes of models and types of modeling applications used by the sponsoring agency were developed.[41] Each scenario consisted of a model, an application for the model, and modifications to the model for the application. Taken together, the scenarios were designed to exercise all parts of the QQR method.

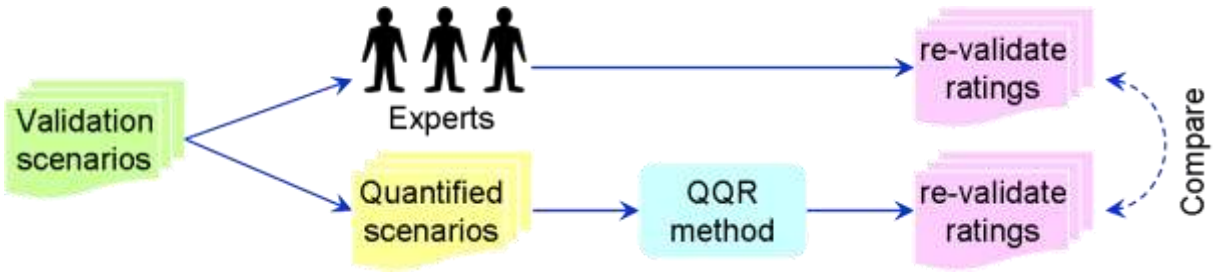


*Figure 5. QQR method validation process.*

[40] Note that there are two distinct "validations" in this section. The first is the re-validation of the modified model, a recommendation for which the QQR method is intended to calculate. The second is the validation of the QQR method itself. The former will always be referred herein as "model re-validation" or similar and the latter as "method validation" or similar.
[41] The validation scenarios are notional. They were developed by UAHuntsville based on public domain information.

| Scenario | Summary |
|---|---|
| 1 | Background: Messages of interest potentially encrypted in pharmaceutical spam<br>Decision: Model used to determine if decryption can be completed in time<br>Variant 1a: Spammer not linked to known threat<br>Variant 1b: Spammer associated with organized crime cartels |
| 2 | Background: Deep web sites found with extensive archives of steganographic photographs<br>Decision: Model used to determine and test download strategy<br>Variant 2a: Perceived threat from information in images is low<br>Variant 2b: Perceived threat from information in images is high |
| 3 | Background: Wifi coverage installation enables surge of access to industrial espionage sites<br>Decision: Model used to determine if wifi coverage will be suspended<br>Variant 3a: Wifi coverage provided by single service provider<br>Variant 3b: Wifi coverage provided by multiple service providers |
| 4 | Background: Expansion of server facility needed to process growing traffic in African nation<br>Decision: Model used to analyze benefits of server facility upgrade<br>Variant 4a: Terrorist groups based in the African nation have made recent attacks<br>Variant 4b: Terrorist groups based in the African nation have not made recent attacks |
| 5 | Background: Expanded email traffic to/from threat nation detected<br>Decision: Model used to determine if current server can process increased traffic<br>Variant 5a: Threat nation leadership has recently undergone a transition<br>Variant 5b: Threat nation leadership is adversarial but stable |
| 6 | Background: Voice and data traffic in and out of a terrorist location must be acquired<br>Decision: Model used to estimate throughput of proposed communications network<br>Variant 6a: Terrorist is linked to “grass roots” urban protest movement<br>Variant 6b: Terrorist is linked to small but heavily armed anti-government group |
| 7 | Background: Sudden increase in encrypted email traffic out of prestigious university<br>Decision: Model used to determine resources needed for decryption<br>Variant 7a: Email traffic originating from physics research laboratory<br>Variant 7b: Email traffic originating from economics department |
| 8 | Background: Large new CONUS computation facility planned<br>Decision: Model used to compare alternative network architectures<br>Variant 8a: Facility is used primarily for long-term data storage<br>Variant 8b: Facility is used primarily for supplemental surge capacity |

*Table 19. Summary of validation scenario and variants.*

The documentation prepared for each of the validation scenarios described the scenario’s background, the role of the model in the real-world decision to be made, and the modifications that were made to the model. Each of the validation scenarios had two variants, meant to be identical in all attributes except one; the difference in the single attribute between the two variants of a scenario was intended to induce a low-high contrast in one of the terms or factors of the QQR main formula. Descriptions of the variants were part of the scenario documentation. The scenario descriptions were written a level of detail and technical specificity intended to be consistent with the information that might be expected to be available to a project manager considering whether or not to re-validate a model. Summaries of the validation scenarios and variants are given in Table 19. Full text of the scenarios may be found in Appendix C.

*Step (2).* Eight experts were asked to assess the scenario variants and make re-validation recommendations for each one. The experts consulted in step (2) were selected due to their expertise in modeling and simulation and/or their familiarity with the sponsoring agency’s modeling applications. Four of the eight experts (designated E1–E4 in the results to appear later) were employees of the sponsoring agency. The other four experts (designated E5–E8) were

UAHuntsville employees; only two were involved with the project. The experts responded independently to each scenario variant, providing 16 responses each. The experts' responses were given in the form of a Likert scale re-validation recommendation for the scenario variant; the possible responses were: Definitely not, Probably not, Neutral, Probably yes, Definitely yes.

*Step (3).* Inputs to the QQR method corresponding to the validation scenarios were prepared. One of the UAHuntsville researchers (W. Colley) selected the missions-mean evidence items and a different UAHuntsville research (M. Petty) selected the qualitative levels for the other terms and factors. Their selections were chosen to reflect the validation scenarios and variantss. They did not consult with each other while preparing their own inputs, nor did they review the other's inputs. The QQR inputs were prepared before the expert's recommendations were examined; i.e., the experts' recommendations were not known at the time the QQR inputs were prepared.

*Step (4).* The QQR inputs prepared in Step (3), both the missions-means evidence items and the qualitative-to-quantitative terms and factors, were input into a spreadsheet implementing the QQR method calculations. The spreadsheet calculated a re-validation recommendation value for each scenario variant. As an example, Figure 6 shows scenario variant 1b as it was input to the QQR method spreadsheet. The qualitative-to-quantitative terms and factors of the method, and its output for scenario variant 1a as p(*unacceptable* | *modifications*) = 0.3542 can be seen at the bottom of the smaller spreadsheet excerpt. The larger spreadsheet excerpt is the missions-means matrix with the evidence items selected for the scenario variant.

*Step (5).* The experts' re-validation recommendations and the QQR method's recommendations were compared using a statistical measure of correlation. The comparison and its results will be described in some detail in the next section.

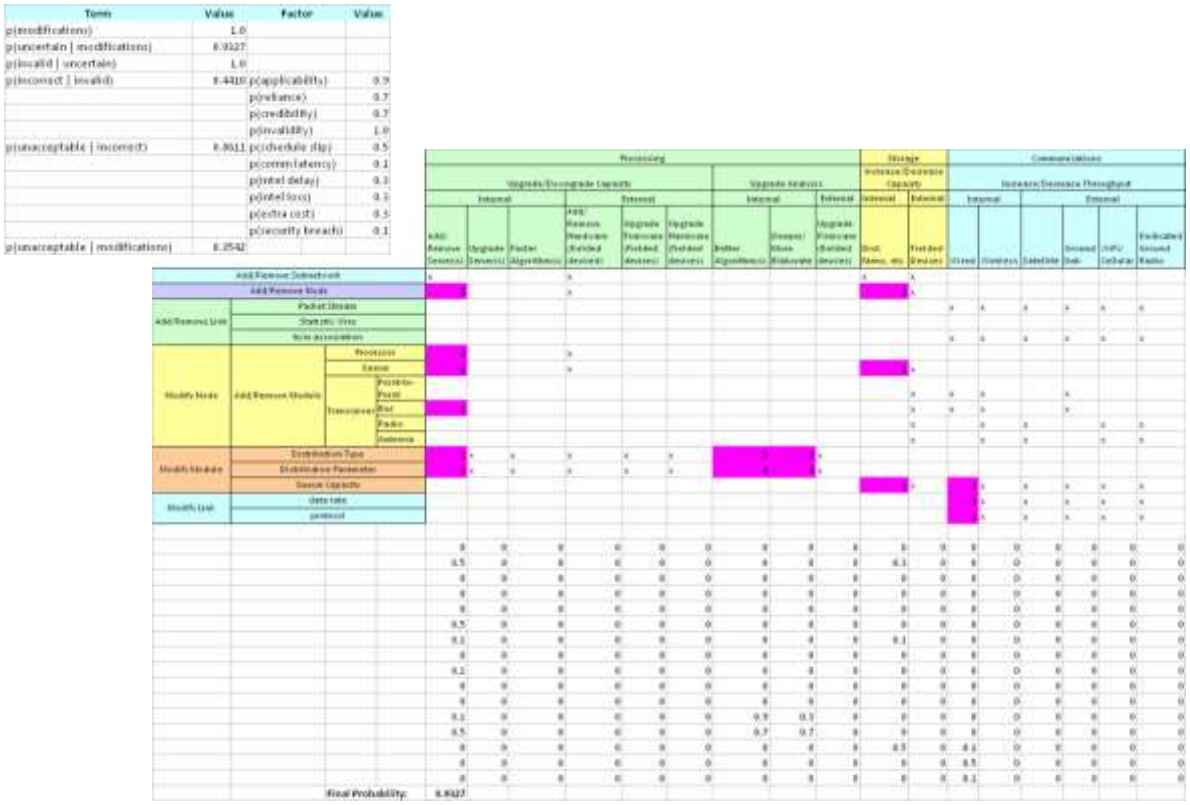

*Figure 6. QQR input example: Scenario variant 1a.*

## 6.2 Validation results

| Scenario variant | Expert re-validation recommendations | | | | | | | | Experts summary statistics | | | | QQR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | Eμ | Eσ | Emin | Emax | |
| 1a | 2 | 3 | 2 | 2 | 1 | 4 | 3 | 3 | 2.50 | 0.93 | 1 | 4 | 0.3542 |
| 1b | 5 | 4 | 5 | 3 | 2 | 5 | 5 | 4 | 4.13 | 1.13 | 2 | 5 | 0.3977 |
| 2a | 1 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2.13 | 0.64 | 1 | 3 | 0.3452 |
| 2b | 3 | 4 | 4 | 3 | 3 | 5 | 4 | 4 | 3.75 | 0.71 | 3 | 5 | 0.4065 |
| 3a | 2 | 3 | 4 | 2 | 3 | 4 | 3 | 3 | 3.00 | 0.76 | 2 | 4 | 0.3052 |
| 3b | 5 | 4 | 4 | 4 | 3 | 4 | 3 | 4 | 3.88 | 0.64 | 3 | 5 | 0.3229 |
| 4a | 4 | 5 | 5 | 4 | 4 | 5 | 5 | 5 | 4.63 | 0.52 | 4 | 5 | 0.5262 |
| 4b | 5 | 4 | 5 | 4 | 3 | 4 | 4 | 4 | 4.13 | 0.64 | 3 | 5 | 0.5224 |
| 5a | 1 | 5 | 5 | 5 | 3 | 5 | 5 | 4 | 4.13 | 1.46 | 1 | 5 | 0.3877 |
| 5b | 2 | 4 | 5 | 5 | 4 | 4 | 2 | 4 | 3.75 | 1.16 | 2 | 5 | 0.3854 |
| 6a | 5 | 5 | 5 | 5 | 4 | 5 | 5 | 4 | 4.75 | 0.46 | 4 | 5 | 0.3986 |
| 6b | 5 | 5 | 5 | 5 | 3 | 5 | 5 | 5 | 4.75 | 0.71 | 3 | 5 | 0.4120 |
| 7a | 3 | 4 | 4 | 3 | 4 | 5 | 2 | 5 | 3.75 | 1.04 | 2 | 5 | 0.3937 |
| 7b | 4 | 4 | 3 | 4 | 3 | 3 | 2 | 3 | 3.25 | 0.71 | 2 | 4 | 0.3788 |
| 8a | 5 | 5 | 4 | 5 | 4 | 5 | 4 | 3 | 4.38 | 0.74 | 3 | 5 | 0.3704 |
| 8b | 4 | 5 | 4 | 5 | 4 | 5 | 4 | 4 | 4.38 | 0.52 | 4 | 5 | 0.4196 |

*Table 20. Summary of experts' re-validation recommendations for the scenario variants.*

Table 20 summarizes the responses given by the experts, shown in columns E1–E8. In the table, the responses are coded as follows: 1 = Definitely not, 2 = Probably not, 3 = Neutral, 4 = Probably yes, and 5 = Definitely yes. The column labeled Eμ is the mean value of the expert responses, Eσ is the standard deviation of the expert responses, Emin is the minimum expert response, and Emax is the maximum expert response.[42,43] The column labeled QQR in the table is the output value for p(*unacceptable* | *modifications*) calculated by the QQR method.

The experts' responses and the QQR method responses were compared statistically using Goodman and Kruskal's gamma statistic.[44] By way of background, Goodman and Kruskal's

[42] Note the extreme range of expert responses on several scenario variants: on 5a, the range is 5; on 1a, 1b, 5b, and 7a, the range is 4.

[43] It must be stated that calculation of the mean and standard deviation from discrete Likert-scale values, as done in the Eμ and Eσ columns, is methodologically questionable. The reason is that it assumes that the experts used the Likert ratings as numerical scalars, e.g., that a "Probably yes" rating, coded as 4, means that a rater was 2× as likely to recommend re-validation as compared to a "Probably not" rating, coded as 2. That assumption is generally unsupportable. Nevertheless, the questionable summary statistics are reported in the table as a simple way to compare the experts' opinions of the scenario variants; they should be interpreted as suggestive only.

[44] The statistic is sometimes denoted γ or Γ instead of gamma.

gamma is a nonparametric statistic used to evaluate the correlation between two variables.[45] While there are other possibly relevant statistics (e.g., Spearman's rho and Kendall's tau), the gamma statistic is preferred when there are multiple ties in the data, i.e., when two or more values for a given variable are equal. Sample applications of the gamma statistic include pharmaceutical clinical trials [Chen, 1999], financial investments [Hryniewicz, 2006], and military modeling and simulation standards [Petty, 2011].

Conceptually, the gamma statistic is evaluated by performing pairwise comparisons. For every possible pair of members of the sample, the rank order of the variables of interest on one member is compared to the rank order of the same variables on the other member. If the rank order of the variables is the same for the two members, they are considered a *concordant* pair and if the rank order is different they form a *discordant* pair. If either of the members have equal values for the two variables, the pair is dropped.[46] The gamma statistic is calculated by subtracting the number of discordant pairs, denoted as *D*, from the number of concordant pairs, denoted as *C*, and then dividing by the sum of *C* and *D*:

$$\Gamma = \frac{C - D}{C + D}$$

The gamma statistic falls within the interval (1, –1), where 1 indicates a perfect positive correlation, –1 indicates a perfect negative correlation, and 0 indicates that there is no correlation and that the variables are independent.

For the QQR method validation, the gamma statistic was calculated in the following manner. For each of the 120 pairs of scenario variants and each of the 8 experts, the inequality relationship (either < or >) of the expert's re-validation recommendations for the pair of scenario variants was compared with the inequality relationship of the QQR method's outputs for the same pair of scenario variants. If the expert's recommendations and the QQR method's outputs stood in the same relationship (either < or > for both the expert and QQR), the pair was counted as concordant; if the relationship was different (either < for the expert and > for QQR, or vice versa), the pair was counted as discordant. If either the expert or QQR assigned the same recommendation to both scenario variants, the pair was considered a tie and not counted in the statistic.

For example, consider expert E4 and scenario variants 7a and 8b in Table 20. Expert E4's response to 7a was "Neutral" (coded as 3) and 8b was "Definitely yes" (coded as 5); expert E4's recommendation for 8b was higher than for 7a. The QQR method output for 7a was 0.3937 and for 8b was 0.4196, thus QQR's recommendation for 8b was also higher than 7a. This is a concordant pair. If the QQR method's ratings had been reversed, this would have been a concordant pair.

Table 21 reports the validation correlation results. Each row in the table shows the number of concordant pairs, discordant pairs, and ties for one of the experts compared to the QQR method; the gamma statistic, which measures correlation, is also given for each expert.

[45] Nonparametric statistics may be used when assumptions can not be made about a population's distribution [Brase, 2009].

[46] The latter situation, known as a "tie", is rather likely in the QQR validation data; e.g., an expert may assign "Probably yes" (coded 4) to two scenario variants.

| Expert | QQR | *C* | *D* | Ties | Γ |
|---|---|---|---|---|---|
| E1 | QQR | 68 | 29 | 23 | 0.4021 |
| E2 | QQR | 69 | 12 | 39 | 0.7037 |
| E3 | QQR | 62 | 21 | 37 | 0.4940 |
| E4 | QQR | 65 | 28 | 27 | 0.3978 |
| E5 | QQR | 58 | 25 | 37 | 0.3976 |
| E6 | QQR | 61 | 13 | 46 | 0.6486 |
| E7 | QQR | 70 | 25 | 25 | 0.4737 |
| E8 | QQR | 67 | 16 | 37 | 0.6145 |
| Total | | 520 | 169 | 271 | 0.5094 |

*Table 21. Validation correlation results.*

The overall value for the correlation statistic was 0.5094, as shown:

$$\Gamma = \frac{C - D}{C + D} = \frac{520 - 169}{520 + 169} \approx 0.5094$$

We interpret this value as strong positive correlation between the experts and the QQR method. Interestingly, observe that the expert with whom QQR agreed most strongly was expert E2, one of the experts working at the sponsoring agency, and the expert with whom QQR agreed least strongly was expert E5, one of the experts working at UAHuntsville.

A close examination of the correlation results for each scenario variant (not shown here) reveals that scenario variant 3b was the most problematic. (See Table 20 for a summary and Appendix C for the full text of scenario variant 3b.) It had the lowest correlation numerator ($C - D$) among all of the scenario variants and was alone responsible for 19 of the 169 discordant pairs.

# 7. Results and future work

This section summarizes the project's findings and identifies potential future work.

## 7.1 Results

A new method (the "QQR" method) was developed to make a recommendation regarding whether or not a modified model should be re-validated. The QQR method is based on a qualitative-to-quantitative assessment of the risk of unacceptable consequences resulting from the use of an incorrect model. The QQR method combines ideas from several existing methods with addition features unique to the QQR method. Its central feature is a missions-means decomposition that allows the method's user to specify the modifications made to the model in an unambiguous manner and supports the method's estimation of the probability of the modifications making the model invalid.

The QQR method was validated by comparing its re-validation recommendations to those of eight human experts; the comparison was done using a correlation statistic chosen to suit the specifics of the QQR validation. The statistic showed strong positive correlation between re-validation recommendations of the experts and the QQR method.

The QQR method was implemented in prototype form as a spreadsheet, which allows the user to easily use the mission-means decomposition and input the qualitative-to-quantitative terms and factors of the method.

## 7.2 Future work

Although the results of the QQR method development and validation are very promising, there are ample opportunities for future work. Possible future work tasks include the following:

1. Implement the QQR method as a robust and well-documented software product, rather than a spreadsheet prototype.
2. Repeat the validation process with both a larger set of validation scenarios and a larger set of human experts.
3. Enhance the validation process with additional validation scenarios that may be expected to be assessed (by the method and by experts) to be less likely to require re-validation.
4. Extend the missions portion of the missions-means decomposition to include types of modeling applications beyond those in the initial project (i.e., other than network computer and communications systems).
5. Extend the means portion of the missions-mean decomposition to include features and capabilities of model classes beyond those in the initial project (i.e., other than discrete event simulation, especially OPNET models).
6. Develop a formalized process for generalizing the QQR method to new types of modeling applications and classes of models (as reflected in the missions-means decomposition) and to new user organizations (as reflected in the factors of the p(*unacceptable* | *incorrect*) term).
7. Improve the "green, yellow, red" output mapping through an exhaustive or a Monte Carlo analysis of the distribution of possible method output values.
8. Conduct a data collection effort to better quantify the p(*invalid* | *uncertain*) term of the QQR main formula.
9. Conduct a data collection effort to better quantify the p(invalidity) factor of the p(*incorrect* | *invalid*) term of the QQR main formula.

## 8. References

[Asbeck, 1984] E. Asbeck, and Y. Y. Haimes, "The partitioned multiobjective risk method", *Large Scale Systems*, Vol. 6, No. 1, 1984, pp. 13-38.

[Balci, 1981] O. Balci and R. G. Sargent, "A Methodology for cost-risk analysis in the statistical validation of simulation models", *Communications of the ACM*, Vol. 27, No. 4, pp. 190-197.

[Balci, 1985] O. Balci and R. E. Nance, "Formulated problem verification as an explicit requirement of model credibility", *SIMULATION*, Vol. 45, No. 2, pp. 76-86.

[Balci, 1996] O. Balci, "Verification, validation, and accreditation", *Proceedings of the 1998 Winter Simulation Conference*, Washington DC, December 13-16 1996, pp. 41-48.

[Balci, 1998a] O. Balci, "Verification, Validation, and Testing", in J. Banks (Editor), *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, John Wiley & Sons, New York NY, 1998, pp. 335-393.

[Balci, 1998b] O. Balci, "Verification, Validation, and Accreditation", *Proceedings of the 1998 Winter Simulation Conference*, Washington DC, December 13-16 1996, pp. 41-48.

[Balci, 1990] O. Balci, "Guidelines for Successful Simulation Studies", *Proceedings of the 1990 Winter Simulation Conference*, New Orleans LA, December 9-12 1990, pp. 25-32.

[Balci, 1991] O. Balci, "Verification, Validation, and Testing of Models", in S. I. Glass and C. M. Harris (Editors), *Encyclopedia of Operations Research and Management Science*, Kluwer Academic Publishers, Boston MA, 1991.

[Balci, 2000] O. Balci, W. F. Ormsby, J. T. Carr, and S. D. Saadi, "Planning for Verification, Validation, and Accreditation of Modeling and Simulation", *Proceedings of the 2000 Winter Simulation Conference*, Orlando FL, December 10-13 2000, pp. 829-839.

[Balci, 2002] O. Balci, R. E. Nance, J. D. Arthur, and W. F. Ormsby, "Expanding Our Horizons in Verification, Validation, and Accreditation Research and Practice", *Proceedings of the 2002 Winter Simulation Conference*, San Diego CA, December 8-11 2002, pp. 653-663.

[Banks, 2005] J. Banks, J. S. Carson, B. L. Nelson, and D. M. Nicol, *Discrete-Event System Simulation, Fourth Edition*, Prentice Hall, Upper Saddle River NJ, 2005.

[Banks, 2010] J. Banks, J. S. Carson, B. L. Nelson, and D. M. Nicol, *Discrete-Event System Simulation, Fifth Edition*, Prentice Hall, Upper Saddle River NJ, 2010.

[Brase, 2009] C. H. Brase and C. P. Brase, *Understandable Statistics: Concepts and Methods, Ninth Edition*, Houghton Mifflin, Boston MA, 2009.

[Chen, 1999] M. Chen and F. Kianifard, "Application of Goodman-Kruskal's Gamma for Ordinal Data in Comparing Several Ordered Treatments: A Different Approach", *Biometrical Journal*, Vol. 41, No. 4 1999, pp. 491-498.

[Chetouane, 2012] F. Chetouane, K. Barker, and A. S. Viacaba Oropeza, "Sensitivity analysis for simulation-based decision making: Application to a hospital emergency service design", *Simulation Modelling Theory and Practice*, Vol. 20, 2012, pp. 99-111.

[Conwell, 2000] C. L. Conwell, R. Enright, and M. A. Stutzman, "Capability Maturity Models Support of Modeling and Simulation Verification, Validation, and Accreditation", in *Proceedings of the 2000 Winter Simulation Conference*, Orlando FL, December 10-13 2000, pp. 819-828.

[DOD, 2009] Department of Defense, *Instruction 5000.61*, M&S VV&A, 2009.

[Drake, 1992] F. Drake and D. Sobel, *Is Anyone Out There? The Scientific Search for Extraterrestrial Intelligence*, Delacorte, New York NY, 1992 pp. 55-62.

[Elele, 2007] J. N. Elele, *A Methodology for Assessing and Prioritizing the Risks Associated with the Level of Verification, Validation and Accreditation of Models and Simulations*, Presentation, November 12-15 2007.

[Epp, 2011] S. S. Epp, *Discrete Mathematics with Applications, Fourth Edition*, Brooks/Cole, Boston MA, 2011.

[Fontaine, 2009] M. D. Fontaine, D. P. Cook, C. D. Combs, J. A. Sokolowski, and C. M. Banks, "Modeling and Simulation: Real-World Examples", in J. A. Sokolowski and C. M. Banks (Editors), *Principles of Modeling and Simulation: A Multidisciplinary Approach*, John Wiley & Sons, Hoboken NJ, 2009, pp. 181-245.

[Ford, 1997] T. Ford, "Helicopter simulation", *Aircraft Engineering and Aerospace Technology*, Vol. 69, No. 5, 1997, pp. 423-427.

[Gehman, 2003] H. W. Gehman, et al, *Columbia Accident Investigation Board Report Volume I*, National Aeronautics and Space Administration, August 2003.

[Gersting, 2007] J. L. Gersting, *Mathematical Structures for Computer Science, A Modern Treatment of Discrete Mathematics, Sixth Edition*, W. H. Freeman, New York NY, 2007.

[Hale, 2007] J. P. Hale, B. L. Hartway, and D. A. Thomas, "A Common M&S Credibility Criteria-set Supports Multiple Problem Domains", *Presentation at the 5th Joint Army-Navy-NASA-Air Force (JANNAF) Modeling and Simulation Subcommittee Meeting*, May 14-17 2007.

[Harmon, 2005] S. Y. Harmon and S. M. Youngblood, "A Proposed Model for Simulation Validation Process Maturity", *Journal of Defense Modeling and Simulation*, Vol. 2, No. 4, October 2005, pp. 179-190.

[Hartley, 2010] D. Hartley and S. Starr, in A. Kott and G. Citrenbaum (Editors), *Estimating Impact: A Handbook of Computational Methods and Models for Anticipating Economic, Social, Political and Security Effects in International Interventions*, Springer, New York NY, 2010, pp. 311-336.

[Hryniewicz, 2006] O. Hryniewicz, "Goodman-Kruskal γ measure of dependence for fuzzy ordered categorical data", *Computational Statistics & Data Analysis*, Vol. 51, 2006, pp. 323-334.

[IEEE, 2011] Institute of Electrical and Electronics Engineers, *Draft Standard for System and Software Verification and Validation*, IEEE P1012™/D2.0 with editorial corrections, December 2011.

[JHU APL, 2011] Johns Hopkins University Applied Physics Laboratory, *Risk Based Methodology for Verification, Validation, and Accreditation (VV&A), M&S Use Risk Methodology (MURM)*, Technical Report NSAD-R-2011-011, April 2011.

[Kesserwan, 1999] N. Kesserwan, *Flight Simulation*, M.S. Thesis, McGill University, Montreal Canada, 1999.

[Kott, 2010] A. Kott and G. Citrenbaum (Editors), *Estimating Impact: A Handbook of Computational Methods and Models for Anticipating Economic, Social, Political and Security Effects in International Interventions*, Springer, New York NY, 2010.

[Likert, 1932] R. Likert, "A Technique for the Measurement of Attitudes", *Archives of Psychology*, Vol. 140, 1932, pp. 1–55.

[Liu, 2005] F. Liu and M. Yang, "Validation of System Models", Proceedings of the IEEE International Conference on Mechatronics and Automation, Niagara Falls Canada, July 29-August 1 2005, pp. 1721-1725.

[NASA, 2008] National Aeronautics and Space Administration, *Standard for Models and Simulation*, NASA-STD-7009, 7-11-2008.

[Neal, 2005] K. D. Neal, W. N. Colley, and M. D. Petty, "An Application and Evaluation of a Methodology for Testing Operational Level Training Simulations", *Proceedings of the Fall 2005 Simulation Interoperability Workshop*, Orlando FL, September 18-23 2005, pp. 271-279

[Petty, 2010] M. D. Petty, "Verification, Validation, and Accreditation", in J. A. Sokolowski and C. M. Banks (Editors), *Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains*, John Wiley & Sons, Hoboken NJ, 2010, pp. 325-372.

[Petty, 2011] M. D. Petty, R. L. Dunning, and A. J. Collins, "Correlation of Characterizing Attributes and Success in Military M&S Standards", *Proceedings of the Fall 2011 Simulation Interoperability Workshop*, Orlando FL, September 19-23 2011, pp. 188-200.

[Reynolds, 2009] P. F. Reynolds, "The Role of Modeling and Simulation", in J. A. Sokolowski and C. M. Banks (Editors), *Principles of Modeling and Simulation: A Multidisciplinary Approach*, John Wiley & Sons, Hoboken NJ, 2009, pp. 25-43.

[Salmon, 2009] F. Salmon, "Recipe for Disaster: The Formula That Killed Wall Street", *Wired*, Vol. 17, No. 3, March 2009.

[Sargent, 2000] R. G. Sargent, "Verification, Validation, and Accreditation of Simulation Models", *Proceedings of the 2000 Winter Simulation Conference*, Orlando FL, December 10-13 2000, pp. 50-59.

[Sheehan, 2003] J. H. Sheehan, P. H. Deitz, B. E. Bray, B. A. Harris, and A. B. H. Wong, "The Military Missions and Means Framework", *Proceedings of the 2003 Interservice/Industry Training, Simulation, and Education Conference*, Orlando FL, December 1-4 2003, pp. 648-661.

[Tolk, 2012] A. Tolk, "Verification and Validation", in A. Tolk (Editor), *Engineering Principles of Combat Modeling and Distributed Simulation*, John Wiley & Sons, Hoboken NJ, 2012, pp. 263-294.

[Youngblood, 2000] S. M. Youngblood, D. K. Pace, P. L. Eirich, D. M. Gregg, and J. E. Coolahan, "Simulation Verification, Validation, and Accreditation", in *Johns Hopkins APL Technical Digest*, Vol. 21, No. 3, 2000, pp. 359-367.

[Youngblood, 2003] S. M. Youngblood, "Literature Review and Commentary on the Verification, Validation, and Accreditation of Models", *Proceedings of the 1993 Summer Computer Simulation Conference*, Boston MA, July 19-21 1993, pp. 10-17.

## 9. Authors' biographies

**Philip W. Alldredge** is a Research Scientist I in the University of Alabama in Huntsville's Rotorcraft Systems Engineering and Simulation Center. He earned an M.S. in Computer Science from the University of Alabama in Huntsville in 2010. His research interests include model-driven engineering, embedded systems, cloud computing.

**J. Cameron Beach** is a Student Research Specialist in the University of Alabama in Huntsville's Rotorcraft Systems Engineering and Simulation Center and an undergraduate student at the University of Alabama in Huntsville, majoring in Industrial and Systems Engineering. He expects to graduate in May 2013. His research interests include systems integration, statistical analysis, and human factors.

**Wesley N. Colley** is a Senior Research Scientist at the University of Alabama in Huntsville's Center for Modeling, Simulation, and Analysis. He earned a Ph.D. in Astrophysics from Princeton University in 1998. He teaches statistics for the UAHuntsville College of Business and continuous system simulation for the university's Modeling and Simulation graduate degree program. His research interests include physics-based modeling, numerical methods, and model validation methods.

Dr. Colley has spent most of the last decade conducting research in modeling and simulation across a wide array of subject areas, including verification, validation and accreditation (VV&A); discrete event simulation; test and evaluation; cost modeling; supply chain modeling; and statistical analysis. Colley's most significant VV&A activities have included the development (with Petty) of a missions-means based methodology for testing operational level training simulations for the U. S. Navy COMOPTEVFOR, verification of queuing theory-based models of network throughput for Air Force Space Command, and the development of a novel statistical technique for validating radar and EO/IR signatures of virtual targets for the Army Aviation & Missile Research Development and Engineering Center. Colley has also been involved with several discrete event simulation efforts: development of a highly flexible multi-modal transportation model (Federal Transit Authority, Economic Development Agency), discrete event simulation model of the Army's supply chain to Korea (Army Aviation and Missile Command), and modeling network traffic in a Joint mission thread from the InterTEC Joint Fires Scenario (DoD Test Resource Management Center).

**Mikel D. Petty** is Director of the University of Alabama in Huntsville's Center for Modeling, Simulation, and Analysis. He is also an Associate Professor of Computer Science and a Research Professor of Industrial and Systems Engineering and Engineering Management. Prior to joining UAH, he was Chief Scientist at Old Dominion University's Virginia Modeling, Analysis, and Simulation Center and Assistant Director at the University of Central Florida's Institute for Simulation and Training. He received a Ph.D. in Computer Science from the University of Central Florida in 1997. Dr. Petty has worked in modeling and simulation research and development since 1990 in areas that include verification and validation methods, simulation interoperability and composability, human behavior modeling, multi-resolution simulation, and applications of theory to simulation. He has published over 175 research papers and has been awarded over $14 million in research funding. He served on a National Research Council committee on modeling and simulation, is a Certified Modeling and Simulation Professional, and is an editor of the journal *SIMULATION*. While at Old Dominion University he was the

dissertation advisor to the first and third students in the world to receive Ph.D.s in Modeling and Simulation and is currently coordinator of the M&S degree program at UAHuntsville.

Dr. Petty has led and worked on numerous projects involving verification and validation of simulation models over a period of fifteen years. Selected examples are described. In 1993 he used statistical techniques to validate an algorithm to generate ground reconnaissance routes. Routes planned by the algorithm and by military officers for three terrain areas were compared using a non-parametric statistical hypothesis test to determine if the algorithm performed at a level realistically comparable to the military officers. In 2003 he led a project that used structured face validation to independently validate a military joint logistics transportation and feasibility model. Logistics subject matter experts were recruited to exercise the model using scenarios designed to cover its capabilities and then evaluate it using assessment criteria prepared in advance. The results were tabulated and analyzed. In 2004 he and his colleagues developed a methodology for performing operational test of a simulation for training joint task force commanders and their staffs. The methodology was based on a missions-means decomposition, wherein military tasks to be trained were cross referenced with simulation capabilities. The methodology was successfully tested at a military training exercise. In 2005, he designed a retrodictive validation methodology for use in validating a model of crowd behavior that produces realistic behavior for crowds involved in military actions. Reference scenarios, which were detailed recreations of historical events, were developed and executed, and their results in the model compared with historical outcomes. In 2006, he led a project that used model comparison to conduct independent verification of a queuing theory-based model of headquarters process and response sequences. New Monte Carlo models of scenarios designed to cover the subject model's capabilities were developed and executed. Their results compared to the model's results using statistical confidence interval analysis. In 2011, he used model comparison combined with a statistical hypothesis test to compare and validate three constructive combat models. In addition to the projects, he has published numerous journal and conference papers on subjects related to verification and validation methods, including two tutorial chapters on verification and validation in modeling and simulation textbooks published in 2009 and 2010. As of 2012, he had taught three multi-day short courses on verification and validation methods for government and academic audiences and given eight conference tutorials on the same subject.

## 10. Appendix A: Explanation of the subset assumptions in the QQR main formula

The QQR main formula is a simple product of conditional probabilities. Note that these are not products of probabilities for completely different events (or sets of conditions), in which case the product approach would indicate independence of those events. Instead, the chain of conditional probabilities is motivated by an approach akin to the Drake Equation, famed in astronomy for estimating the number of communicating extraterrestrial civilizations in the Milky Way Galaxy [Drake, 1992]. In that product, one finds a series of probability factors, each contingent on the last; for example, here is a portion of the product: (fraction of planets that develop life) × (fraction of life-bearing planets that produce intelligent life) × (fraction of intelligent civilizations that conduct extraterrestrial communication). Similarly, the QQR formula is a product of probability factors, each contingent on the last; for example, (probability of uncertainty, given modifications) × (probability of invalidity, given uncertainty) × (probability of incorrect usage, given invalidity). In the Drake Equation, it is self-evident that there cannot be an intelligent civilization on a planet without life. In our QQR equation, the contingencies are not self-evident, but assumed. As such, let us therefore examine for each factor the implications of its contingency.

p(*modifications*): This factor is taken to be 1, since the premise of the calculation is that the model has been modified.

p(*uncertain* | *modifications*): The assumption that uncertainty must arise from modification is essentially the assumption that the model was valid before the modification. Put a different way, the QQR treatment only addresses *new* validity uncertainties addressed by the modifications to the model, including modifications that use pre-existing model components in new ways.

p(*invalid* | *uncertain*): As discussed earlier, this factor is taken to be 1, because fundamentally one cannot predict validity in uncertain circumstances.

p(*incorrect* | *invalid*): This factor requires the most care in the present discussion. It presumes that an incorrect real-world decision is made because a model is invalid (Type II errors). Of course, incorrect decisions can be made when one fails to use valid models properly (Type I errors). The QQR method cannot access Type I errors, since the extent of its purpose ends with a recommendation to validate, and does not address whether stake-holders choose to trust a valid model.

p(*unacceptable* | *incorrect*): Here we assume that an unacceptable outcome can only arise from an incorrect decision. This is not strictly true; unacceptable outcomes sometimes arise even when the best decisions are made. However, our concern is valid model support of decisions. If a correct decision is made, then the model has done its job validly (or at least not prevented a correct decision).

Therefore, each factor is conditioned on the last. In that case, the final output probability is not simply p(*unacceptable* | *modifications*); it is really short-hand for

$$p(\textit{unacceptable} \mid \textit{modifications}) = p(\textit{unacceptable} \mid \textit{incorrect} \mid \textit{invalid} \mid \textit{uncertain} \mid \textit{modifications}).$$

Let us now derive our main formula. Recall the multiplication law for conditional probabilities:

$$p(A \cap B) = p(A \mid B) \cdot p(B)$$

Because we have asserted that modifications have been made, p(*modifications*) = 1, and so the left hand side can be re-written as p(*unacceptable* ∩ *modifications*).

Using the multiplication law, we shall construct the right hand side term by term.

p(*unacceptable* ∩ *incorrect*) = p(*unacceptable* | *incorrect*) · p(*incorrect*)

But we are studying only the incorrect decisions that arise from invalidity, so p(*incorrect*) ≡ p(*incorrect* ∩ *invalid*), which is to say *incorrect* ⊂ *invalid*, and that p(*incorrect* | *invalid*) = p(*incorrect*) / p(*invalid*).

p(*unacceptable* ∩ *incorrect*) = p(*unacceptable* | *incorrect*) · p(*incorrect* | *invalid*) · p(*invalid*)

But we are studying only the incorrect decisions that arise from invalidity, so p(*invalid*) ≡ p(*invalid* ∩ *uncertain*) , which is to say *invalid* ⊂ *uncertain*, and that p(*invalid* | *uncertain*) = p(*invalid*) / p(*uncertain*).

p(*unacceptable* ∩ *incorrect* ∩ *invalid* ∩ *uncertain*) =
p(*unacceptable* | *incorrect*) · p(*incorrect* | *invalid*) · p(*invalid* | *uncertain*) · p(*uncertain*).

But we are studying only the uncertainties that arise from modifications, so p(*uncertain*) ≡ p(*uncertain* ∩ *modifications*) , which is to say *uncertain* ⊂ *modifications*, and that p(*uncertain* | *modifications*) = p(*uncertain*) / p(*modifications*).

p(*unacceptable* ∩ *incorrect* ∩ *invalid* ∩ *uncertain*) =
p(*unacceptable* | *incorrect*) · p(*incorrect* | *invalid*) · p(*invalid* | *uncertain*) ·
p(*uncertain* | *modifications*) · p(*modifications*).

Our presumption is that *p*(*modifications*) = 1, which obviates the final term, and we recover

p(*unacceptable* ∩ *incorrect* ∩ *invalid* ∩ *uncertain* ∩ *modifications*) =
p(*unacceptable* | *incorrect*) · p(*incorrect* | *invalid*) · p(*invalid* | *uncertain*) ·
p(*uncertain* | *modifications*).

Recalling the fact that the left-hand side can be written as simply p(*unacceptable* | *modification*), we recover our original equation,

p(*unacceptable* | *modifications*) =
p(*unacceptable* | *incorrect*) · p(*incorrect* | *invalid*) · p(*invalid* | *uncertain*) ·
p(*uncertain* | *modifications*).

To summarize the above calculation, at each step one makes the identification at each step of p(*B*) ≡ p(*B* ∩ *A*), which is to say that *B* pre-supposes *A* (or that *B* ⊂ *A*), and that p(*B* | *A*) = p(*B*) / p(*A*). This allows a straightforward multiplication of all the conditional probabilities.

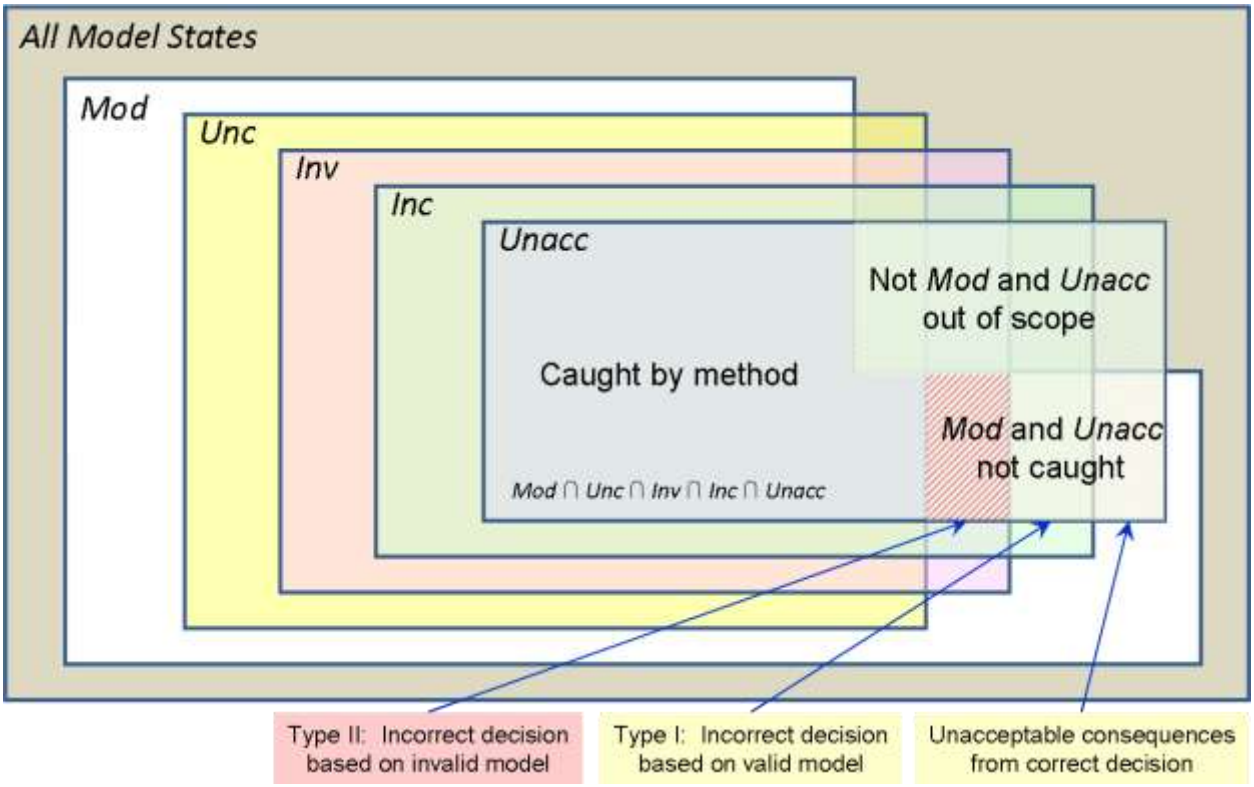


*Figure 7. QQR method scope and assumptions.*

Figure 7 illustrates the situation. The light blue "Caught by method" region is precisely the event the QQR method studies (*unacceptable* ∩ *incorrect* ∩ *invalid* ∩ *uncertain* ∩ *modifications*). Illustrated also are three important areas not caught by the method. First, we have a missed Type II error (red-hashed region), which this method would normally catch; however in this case, the modifications to the model introduced no uncertainties, so re-validation was not required; this condition is quite unlikely since nearly any model modification would introduce some uncertainty. The Type I error region is shown as the left-hand portion of the yellow-hashed region—a valid model was misused in making an incorrect decision that had unacceptable consequences. This is out of our scope since our method can only recommend validation, not whether stake-holders trust the outputs of a valid model. The right half of the yellow-hashed region is where the model is valid and a correct decision is made, but unacceptable consequences still arise, clearly out of our scope. The green-hashed region shows where unacceptable consequences arise with no model modifications. Since the purpose of this method is to consider the need to conduct validation after model modifications, this region is also out of scope.

## 11. Appendix B: Validation scenarios

The complete texts of the scenarios used in the validation process, as described earlier in the Validation section, are given in this appendix. The scenarios are presented here as they were presented to the human experts, including the scenario variants and the Likert scale inputs.

The validation scenarios are notional. They were created entirely by UAHuntsville based on unclassified public domain sources, not by any agency or employee of the U. S. Department of Defense. They not intended to describe, nor should they be interpreted as describing, any actual activities, current or past, of any agency of the U. S. Department of Defense. If there is any resemblance of the activities described in these scenarios to any actual activities of any agency of the U. S. Department of Defense, it is coincidental.

**Scenario 1**

*Background.* Exploratory checking by a Government organization unexpectedly detects possible encrypted messages embedded in recent pharmaceutical sales spam email messages sent by an Eastern European-based spam organization that has been active for several years. Breaking the encryption will require retrieval and analysis of an extremely large amount of archived spam messages sent by the spam organization over the last three years. The Government organization must locate and retrieve the archived messages from file servers in a low-security long-term storage facility, transfer them via a network to a high-security analysis facility, and execute extensive decryption computation. Federal regulations will be cited to impose a deadline for deciding whether to shut down the spam organization via commercial anti-spamming actions or to covertly monitor it as a potential source of intelligence information.

*Model role and decision.* A model of the data network to be used for the transfer and the computational facility to be used for the decryption exists, but it has not been modified to reflect recent hardware upgrades to those systems. Once modified to reflect the upgrades, it will be used to determine whether the retrieval and decryption can be completed in time.

*Modifications to the model.* Parameters representing the data capacity of the links in the data network were changed. Parameters representing the number of servers and the mean service time for the computational facility were changed. Additional probability distributions modeling the time required to find archived messages in the storage facility and the number of relevant archived messages likely to be found were added.

*Scenario 1 variant a.* The spam organization is not known to be linked to any national security threat.

Re-VV&A recommendation for scenario 1a

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

*Scenario 1 variant b.* The spam organization is associated with international organized crime cartels expanding into the U.S.

Re-VV&A recommendation for scenario 1b

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

**Scenario 2**

*Background.* Routine web surveillance by a Government organization discovers several previously unknown "deep web" sites (without URLs or DNS services) containing very large numbers of suspiciously uninteresting photographs. Steganographic encoding of potentially important information in a subset of the images is suspected. The images must be downloaded and analyzed to determine their content and source. The Government organization must decide whether to download the images individually at intervals over a long period of time so as to avoid raising suspicion, or to download the images en masse all at once in the hope of retrieving as many of the images as possible before the accesses are detected and blocked.

*Model role and decision.* A generic model of internet communications and data downloads is available. It will be customized to reflect the specific known or conjectured performance of the Government organization's download servers, the internet communications link to the "deep web" image servers, and the performance of those image servers. Once modified, model will be used to select one of the two download approaches under consideration.

*Modifications to the model.* Parameters representing the data capacity of the links in the internet were changed. Parameters representing the number of the Government organization's download servers and the mean service time for their download operation were changed. Additional probability distributions modeling the time required for the "deep web" server to access and download a requested image and the probability of that server determining that an access should be blocked were added.

*Scenario 2 variant a.* Based on associated intelligence, the perceived threat level from the information concealed in the images is very low to low.

Re-VV&A recommendation for scenario 2a

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

*Scenario 2 variant b.* Based on associated intelligence, the perceived threat level from the information concealed in the images is high to very high.

Re-VV&A recommendation for scenario 2b

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

**Scenario 3**

*Background.* The installation and activation of full wifi coverage of an entire small town in rural Oregon is followed by an expected spike in accesses and data transfers to and from web sites. However, a surprisingly large amount of the new traffic involves web sites associated with foreign commercial companies that have been implicated in attempts to collect information about U.S. military science and technology development. Further analysis of both the sources and destinations of the traffic is needed to determine if the accesses represent a threat or if they are within the range of expected variation.

*Model role and decision.* A general-purpose model of internet activity based on network theory is available. The model must be modified and parameterized to represent the Oregon town's demographics. Based on the model's estimation of the probability of the observed internet activity, a decision will be made as to whether the wifi coverage should be suspended or discontinued.

*Modifications to the model.* Parameters describing the distribution and frequency of sites internet sites visited based on user demographics were added to the internet activity model.

*Scenario 3 variant a.* The town's wifi access is routed through a single internet service provider.

Re-VV&A recommendation for scenario 3a

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

*Scenario 3 variant b.* The town's wifi access is supported by multiple internet service providers.

Re-VV&A recommendation for scenario 3b

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

**Scenario 4**

*Background.* An African nation has seen dramatically increased communications capabilities over the last year, which has increased communications volume by a factor of two. The extra traffic is starting to burden general servers at a U.S. Government facility. The nation is not overtly anti-American, but the desperately poor government cannot prevent the nation from being a haven to Islamic terrorists. The U.S. Government facility creates a small, dedicated server area and re-directs traffic to and from this nation to the new servers.

*Model role and decision.* A model is to be used to determine benefits of new server area in terms of increased storage and processing for the new threat situation, and in terms of restored throughput/storage on the original (general) servers.

*Modifications to the model.* Introduction of a new sub-network modeling the dedicated server area. Also, new communications links between principal incoming data link (used for foreign to US traffic) and new sub-network.

*Scenario 4 variant a.* Terrorist groups in this nation have recently taken credit for instigating an assault against a US embassy. Increased communications infrastructure and traffic is primarily e-mail.

Re-VV&A recommendation for scenario 4a

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

*Scenario 4 variant b.* Terrorist groups in this nation are only loosely organized and not thought to be immediate threats. Increased communications infrastructure and traffic is primarily cellular voice and text.

Re-VV&A recommendation for scenario 4b

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

**Scenario 5**

*Background.* A significant increase is observed in e-mail traffic between a threat nation and U. S. citizens. More intensive monitoring and analysis with respect to this nation's data traffic in and out of the U.S. is deemed necessary. Existing network, storage and analysis resources are prioritized for this nation's traffic in and out of U.S.

*Model role and decision.* A model is to be used to determine if the new loads compromise uptime and utility of server. The model must also handle the different geographic distributions of the data pulls in the different sites.

*Modifications to the model*. Loads on links, storage and processing on the existing servers and network increase. New channels of data from external links into a U.S. Government facility from the threat nation monitoring sites are created.

*Scenario 5 variant a.* The threat nation has recently undergone a transition in power. The posture of the new regime toward U.S. is still unclear. E-mail is traveling to a geographically scattered set of citizens inside CONUS.

Re-VV&A recommendation for scenario 5a

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

*Scenario 5 variant b.* Threat nation is a stable, though adversarial, dictatorship. E-mail traffic is mainly concentrated into Bay Area of California.

Re-VV&A recommendation for scenario 5b

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

**Scenario 6**

*Background.* A major cellular provider grants access to voice and data traffic in and out of suspected domestic terrorist group. A secure hardwired link is set up between one of the provider's data sites and a SIPRNet location at a nearby military installation. New data servers to handle the large dump of data are created. Analysis processing capability is assigned to the dataset.

*Model role and decision.* Model must determine the throughput of the arrangement all the way back to a U.S. Government facility's new data servers. Model is also used to determine the processing time required for analyzing the new dataset.

*Modifications to the model.* Introduction of the new link between the provider and the SIPRNet location. Introduction of "effective" sub-network of military installation's particular SIPRNet configuration. New data servers must be introduced to model. Existing processors are re-assigned to the new data, and therefore their location within the network topology is changed. Model must also take into consideration the rate at which the data can be pulled by cellular provider from its network and provided to the SIPRNet site.

*Scenario 6 variant a.* Terrorist group is linked by FBI to "grass-roots" urban protest movement seen across dozens of major cities. Communication is primarily voice/text.

Re-VV&A recommendation for scenario 6a

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

*Scenario 6 variant b.* Terrorist group is linked with a small, remote, though heavily armed anti-government group. A significant fraction of cellular traffic appears to be encrypted data transfers.

Re-VV&A recommendation for scenario 6b

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

**Scenario 7**

*Background.* A U.S. Government organization detects a sudden increase in large encrypted e-mails flowing out of University of Chicago to a foreign entity. The encryption is custom (non-RSA) and appears quite strong. The U.S. Government organization determines that substantial cryptography assets should be assigned to this issue.

*Model role and decision.* Model is used to determine the impact of the new servers on the problem at hand. Model is also used to determine backlogs in general cryptography processing introduced by the redirection of assets.

*Modifications to the model.* A significant number of processing servers are moved within the network. A dedicated data server for just these e-mails is connected directly to the processing servers for optimal performance.

*Scenario 7 variant a.* The e-mail is determined to be originating from the campus offices of Fermilab (Department of Energy) employees.

Re-VV&A recommendation for scenario 7a

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

*Scenario 7 variant b.* The e-mail is from the University of Chicago Economics department, and encryption scheme is found to require a time-sensitive pass-phrase. The time-sensitive pass-phrase requires major processing assets; however the economics department origins are regarded to be somewhat less threatening than other possible UC locations.

Re-VV&A recommendation for scenario 7b

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

**Scenario 8**

Unlike the other scenarios, in this scenario the model role and decisions and model modifications depend on the scenario variant.

*Background.* A major new CONUS computation facility for a U. S. government organization is planned. The new facility will handle information under a broad domain, such as all information originating from or going to threat entities within allied nations.

*Scenario 8 variant a.* The new site is primarily a long-term data storage facility for archived information, with initial storage and processing still occurring at its original location. New dedicated fiber connections are created between the facilities.

*Model role and decision for scenario 8a.* Model is to be used to determine the benefits of unburdening the data servers at the original location; also model will be used to estimate latencies introduced when processing requires data pulls back from the new facility.

*Modifications to the model for scenario 8a.* Major storage model must be constructed to represent the new facility. Major links must be created to represent the new fiber connections.

Re-VV&A recommendation for scenario 8a

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

*Scenario 8 variant b.* The new site operates primarily as a surge/critical processing capacity facility, rather than as a long-term general-purpose processing and storage site. During demand surges and critical situations, data streams are routed directly to the new facility, though data and processing artifacts will be routed back to the original location over time. Nonetheless, dedicated fiber connections are created between the facilities.

*Model role and decision for variant 8b.* Model is to be used to determine the efficacy of the approach, in terms of trigger points for re-routing information to the new facility, and estimates of increase in processing power after the trigger is tripped.

*Modifications to the model for scenario 8b.* Major processing facility and surrounding network must be constructed to represent the new facility. Major links must be created to represent the new fiber connections. Significant modifications to models of the existing original location's infrastructure are necessary since surge traffic will be routed directly to the new facility.

Re-VV&A recommendation for scenario 8b

Definitely not ___ Probably not ___ Neutral ___ Probably yes ___ Definitely yes ___

## End of Document